Comparative genome analysis of *Streptococcus pneumoniae* and its close relatives

Vom Fachbereich Biologie der Technischen Universität Kaiserslautern
zur Verleihung des akademischen
Grades “Doktor der Naturwissenschaften” genehmigte

DISSERTATION

von

Dipl.-Bioinform. (FH) Martin Rieger

Datum der wissenschaftlichen Aussprache: 04.09.2020

Vorsitzender der Prüfungskommission: Herr Prof. Dr. Matthias Hahn
Erster Berichterstatter: Frau Prof. Dr. Regine Hakenbeck
Zweiter Berichterstatter: Herr Prof. Dr. John Cullum

Lehrbereich Mikrobiologie der Technischen Universität Kaiserslautern

Kaiserslautern, 2020

D386

PUBLISHED SCIENTIFIC PAPERS

The cumulative thesis is based on the following five publications. The articles in this thesis are identical to the published versions.

Rieger M, Mauch H, Hakenbeck R. **Long persistence of a novel *Streptococcus pneumoniae* 23F clone in a cystic fibrosis patient.** mSphere. 2017 Jun 7;2(3). pii: e00201-17. doi: 10.1128/mSphere.00201-17. eCollection 2017 May-Jun.

Rieger M, Denapaite D, Brückner R, Maurer M, Hakenbeck R. **Draft genome sequences of two *Streptococcus pneumoniae* serotype 19A sequence type 226 clinical isolates from Hungary, Hu17 with high-level beta-lactam resistance and Hu15 of a penicillin-sensitive phenotype.** Genome Announc. 2017 May 18;5(20). pii: e00401-17. doi: 10.1128/genomeA.00401-17.

Denapaite D, **Rieger M**, Köndgen S, Brückner R, Ochigava I, Kappeler P, Mätz-Rensing K, Leendertz F, Hakenbeck R. **Highly variable *Streptococcus oralis* strains are common among viridans *Streptococci* isolated from primates**. mSphere. 2016 Mar 9;1(2). pii: e00041-15. doi: 10.1128/mSphere.00041-15. eCollection 2016 Mar-Apr.

Todorova K, Maurer P, **Rieger M**, Becker T, Bui N K, Gray J, Vollmer W, Hakenbeck R. **Transfer of penicillin resistance from *Streptococcus oralis* to *Streptococcus pneumoniae* identifies *murE* as resistance determinant**. Mol Microbiol. 2015 Sep;97(5):866-80. doi: 10.1111/mmi.13070. Epub 2015 Jun 19.

Tettelin H, Chancey S, Mitchell T, Denapaite D, Schähle Y, **Rieger M**, Hakenbeck R. **Genomics, genetic variation, and regions of differences.** 2015 May. In: ***Streptococcus pneumoniae*: Molecular mechanisms of host-pathogen interactions.** pp 81-107. Eds.: Orihuela C, Hammerschmidt S, Brown J. Academic Press, London.

Table of contents

1 Introduction

1.1 Streptococcus

The genus Streptococcus, member of the lactic acid bacteria within the phylum Firmicutes, comprises several species of spherical Gram-positive bacteria who divide in one axis leading to pairs or chains of cells (Cole, et al., 2008). They are immobile and can commonly be found in several warm-blooded animals including human (Cole, et al., 2008). The genus *Streptococcus* is one of the most invasive groups of bacteria, where 35 of 57 species can cause invasive diseases with *S. pneumoniae*, *S. pyogenes* (group A), *S. agalactiae* (group B) and *S. mutans* being the main cause of streptococcal infections in human (Krzyściak W, 2013; Cole, et al., 2008). Despite the ability to cause severe diseases, most *Streptococcus spp*. are commensals (Cole, et al., 2008; Krzyściak W, 2013). Some streptococci are used for the production of medical, cosmetical, nutraceutical (Liu, et al., 2011) and dairy products (Keogh, 1970; Westerik, et al., 2016; Han, et al., 2016).

The definition of streptococcal species is based on several phenotypic and genotypic properties. The introduction of multilocus sequence typing (MLST) (Enright, et al., 1999) and multilocus sequence analysis (MLSA) (Bishop, et al., 2009) which is based on comparative analysis on concatenated sequences of a set of housekeeping genes has been extremely useful to discriminate between closely related species, and to differentiate clones within a species. Other parameters include the serotype, which is determined by surface antigens like the polysaccharide capsule (Baron, 1996; Geno, et al., 2015), their haemolysis behaviour (Facklam, 2002; Shottmuller) and others (Baron, 1996).

According to these parameters, streptococci initially were classified into several groups. In the era of genome sequencing, this classification is mainly based on the analysis of 16S rRNA (Abranches, et al., 2018; Kilian, et al., 2008; Woese, 2000; Kawamura, et al., 1995) and led to eight groups of streptococci named Mitis, Salivarius, Bovis, Mutans, Anginosus, Sanguinis, Downei and Pyogenes group (Abranches, et al., 2018). The Mitis group contains alpha-haemolytic bacteria which oxidise the iron in haemoglobin by producing hydrogen peroxide, leading to a green or brown colour by the generation of methaemoglobin when grown on blood agar plates (Blake, 1916; Barnard, et al., 1996). This group includes the pathogen *Streptococcus pneumoniae* and closely related commensal species *S. mitis*, *S. pseudopneumoniae* and *S. oralis*. The first available complete genomes of these species

where *S. pneumoniae* R6 (Hoskins, et al., 2001), *S. mitis* B6 (Denapaite, et al., 2010) , *S. oralis* Uo5 (Reichmann, et al., 2011) and *S. pseudopneumoniae* IS7493 (Shahinas, et al., 2011). Members of this group are of main interest in the work presented here. They populate the upper respiratory tract of human as part of the commensal flora and are naturally competent for transformation (Bracco, et al., 1957; Reichmann, et al., 2011). Transformation has been discovered by Avery in *S. pneumoniae*, who determined DNA as the "fundamental unit of the transforming principle" (Avery, et al., 1944).

The Pyogenes group initially contained beta-haemolytic bacteria which destroy red blood cells using the cytotoxins Streptolysin S or O (SLS or SLO) (Bhakdi, et al., 1985; Marmorek, 1895; Todd, 1938), giving rise to a clear zone around their colonies on blood agar plates. While the oxygen-sensitive SLO interacts with cholesterol of the cell membrane of eukaryotic cells (Bhakdi, et al., 1985), the haemolysis mechanism of the oxygen-stable SLS is not completely understood (Carr A, 2001; Molloy, et al., 2015). Beta-haemolytic organisms have been further divided into Lancefield groups according to surface antigens (Lancefield, 1933). Since then, Lancefield classification alone became insufficient for identification of beta-haemolytic strains and was complemented by other methods (Facklam, 2002; Abranches, et al., 2018). The introduction and continuous advancement of second and third generation sequencing technologies facilitate their identification by genotyping.

Gamma-haemolytic organisms cause no haemolysis. However, alpha- and gamma-haemolysis can be difficult to distinguish since the composition of the growth medium can influence the manifestation of alpha-haemolysis (Facklam, 2002).

1.1.1 Streptococcus pneumoniae

One important member of the Mitis group streptococci is the species *S. pneumoniae*, also referred to as the pneumococcus, a major human pathogen (MacLeod, et al., 1956; Engholm, et al., 2017; Hiller, et al., 2018; Drijkoningen, et al., 2014). This organism and its pathogenic potential was described 1881 by Sternberg and Pasteur as *Microbe septicemique du salive* (Pasteur, 1881; Watson, et al., 1993) and *Microbe pasteuri* (Watson, et al., 1993; Sternberg, 1885) and was named 1886 *Pneumococcus* due to its potential to cause lung diseases (Watson, et al., 1993; Fraenkel, 1886b). The *Pneumococcus* was renamed 1920 to *Diplococcus pneumoniae* (Watson, et al., 1993; Winslow, et al., 1920) and is now called *Streptococcus pneumoniae* since 1974 (Watson, et al., 1993; Deibel, et al., 1974). Its natural habitat as a mainly commensal species is the upper respiratory tract (Weiser, et al., 2018), but as pathogen it is able to cause severe diseases like meningitis, pneumonia, otitis media and cardiac dysfunction (Loughran, et al., 2019). In 1999, it was described, that 1.1 million deaths worldwide were caused by infections by pneumococci each year (Klein, 1999). To support the search for new treatment methods, the genome of the strain S. pneumoniae R6 was determined in 2001 as the first streptococcal genome and contains 2.038.615 nucleotides (nt) (Hoskins, et al., 2001). This strain is a derivative of the capsule type 2 strain *S. pneumoniae* R36A (Smith, et al., 1979), which was used by Avery *et al.* 1944 in the classical transformation experiments (Avery, et al., 1944). An updated genome version in comparison with its ancestor *S. pneumoniae* D29 was published in 2007 by Lanie *et al.* (Lanie, et al., 2007) after a re-sequencing (2.038.617 nt). The second published genome is of the virulent strain TIGR4 (Tettelin, et al., 2001). *S. pneumoniae* and both genomes were basic for our understanding of the gene content of. In the current work the updated sequence of *S. pneumoniae* strain R6 was used.

1.1.1.1 *Streptococcus pneumoniae R6 and its genome*

Since the unencapsulated *S. pneumoniae* R6 strain is missing the capsule, a major virulence factor, due to a large deletion in the capsule locus, this strain has become the main standard laboratory strain (Smith, et al., 1979; Iannelli, et al., 1999; Hoskins, et al., 2001). It is a perfect example to demonstrate the effects of horizontal gene transfer between different strains and species. Due to its natural competence, genes and gene fragments of different genomes are distributed all over the whole *S. pneumoniae* R6-genome. Comparison with the genome of *S. pneumoniae* TIGR4 (Tettelin, et al., 2001) showed a gene difference of about 10 % (Brückner, et al., 2004). Compared to other bacterial species, in *S. pneumoniae* a variety of repetitive elements like complete and incomplete insertion sequences (IS), BOX-, RUP- (repeat unit of pneumococcus) and other elements can be found, where homologous recombination frequently occurs (Hoskins, et al., 2001). BOX-elements consist of a combination of boxA-, B- or C-repeats and might be involved in the regulation of genes (Hoskins, et al., 2001; Zhang, et al., 2015; Croucher, et al., 2011) while RUP-elements (repeat unit of pneumococcus) are supposed insertion sequence derivatives with the ability to support genomic rearrangements (Oggioni, et al., 1999; Croucher, et al., 2011). Besides the gene cluster responsible for capsule biosynthesis which is not functional due to a large-scale deletion in *S. pneumoniae* R6, a variety of virulence factors are known in *S. pneumoniae* and are described below. The highly diverse and still expanding genome of *S. pneumoniae* is supposed to be a result of maintaining stability in its current ecological niche within the human host (Donati, et al., 2010; Kilian, et al., 2014). Since the strain *S. pneumoniae* R6 is avirulent, it served as a basis for the analysis of the pathogen *S. pneumoniae* (Hoskins, et al., 2001) and has been used to explore the evolution of antibiotic resistance in transformation experiments using highly resistant closely related streptococcal species as donor (Hakenbeck R, 1998), Todorova (Todorova, et al., 2015). Today several thousand *S. pneumoniae* genomes are sequenced and publicly available.

1.1.1.2 *Horizontal gene transfer and genetic variability*

There are three main mechanisms of DNA transfer observed in prokaryotes (Ravin, 1960; Bakkali, 2013): Conjugation, where DNA is transferred from a donor to an acceptor cell, transduction involving a phage, and transformation, where free DNA is taken up by the recipient (Ravin, 1960; Johnston, et al., 2013). Transformation has a major impact on the genomic makeup, the focus of the current work, and is described in more detail below.

The ability of natural transformation, discovered by Avery 1944 (Avery, et al., 1944), allows *S. pneumoniae* and other bacteria to access exogenous DNA, which is freely available in the environment, or which is obtained by fratricide of non-competent cells (Claverys, et al., 2009). A special case is the recombination of DNA from a different chromosome within the same individual cell (Johnston, et al., 2013). Competence is a temporary state of a cell, in which it is capable of genetic transformation. This state is regulated by a set of proteins which is activated by a signal which varies depending on the species (Claverys, et al., 2009). In *S. pneumoniae*, the state of competence is maintained only during a short period during exponential growth in liquid media. It is regulated via the temporary production of a secreted peptide (competence stimulating peptide, CSP; ComC) whose recognition by a membrane associated receptor (ComD) results in the expression of a complex regulatory network that involves the production of a large set of proteins (Halfmann A, 2011; Laux A, 2015; Ahn, et al., 2014; Claverys, et al., 2009; Salvadori, et al., 2019). *S. pneumoniae* requires dsDNA (double-stranded DNA) for transformation. ssDNA (single-stranded DNA) leads to an about 200-fold decreased transformation activity (Claverys, et al., 2009). During binding, the dsDNA is fragmented into pieces of about 6.000 nt (Claverys, et al., 2009). These fragments are transported into the cell while the strands are separated, and the non-transported strand is degraded and its components are released into the surrounding medium (Claverys, et al., 2009). Many proteins are involved in the uptake of DNA (Claverys, et al., 2009). Since ssDNA shows a decreased transformation activity, the overall transformation activity directly after DNA uptake is also decreased for a certain time. This time span is called eclipse. During eclipse, the ssDNA is bound to a protein SsbB and forms the eclipse-complex. SsbB and several other proteins such as RecA, CoiA, DprA and RadA are required for DNA incorporation into the chromosome by homologous recombination (Ravin, 1960; Pasta, et al., 1999; Claverys, et al., 2009; Bakkali, 2013). Due to its transformability, *S. pneumoniae* has become the paradigm to

study partners and genomic consequences of horizontal gene transfer events (Wyres, et al., 2012).

Horizontal gene transfer involving closely related species enables the recipient organism to gain new properties to adapt to new ecological niches and environmental factors, and the evolution of antibiotic resistance is one prominent example (Levin, et al., 2009). Moreover, it facilitates capsule switch which results in evasion from vaccine treatment (Johnston, et al., 2013). Homologous recombination sites are recognized as mosaic genes, where the integration of foreign DNA leads to sequence blocks that are highly distinct from corresponding sequences in the parental strains (Laible, et al., 1991; Hakenbeck, et al., 2001). Popular examples are penicillin binding proteins (PBP) genes especially PBP2b, PBP1a and PBP2x (Hakenbeck R, 1998; Dowson, et al., 1989; Laible, et al., 1989; Laible, et al., 1991; Coffey, et al., 1991).

In addition, the genomes of Streptococci are permanently altered due to spontaneous mutations (Madigan, et al., 2002), and the genome organization can be changed by movement of mobile elements including transposons (Muñoz-López, et al., 2010) and insertion sequences (IS) (Mahillon, et al., 1998). Transposons consist of transposases responsible for excision depending on flanking sequence repeats and a transposon body containing cargo genes (Muñoz-López, et al., 2010). After excision of a transposon, often flanking transposase genes remain in the genome. IS contain only genes encoding proteins for transposition of sequence and facilitate for example chromosome rearrangements and plasmid integration (Mahillon, et al., 1998). Both structures – transposons and IS - facilitate the movement of DNA within one genome (of an individual cell), but not necessarily between cells (Johnson, et al., 2015). Moreover, prophages and phage remnants are frequent in many streptococci (Brueggemann, et al., 2017).

All these factors contribute to the vast variability and diversity of bacterial genomes (Ravin, 1960) and complicate the calculation of the evolutionary tree (Philippe, et al., 2003) and the origin of genetic features. Thus, these elements mostly are not used in such analyses.

1.1.1.3 *Virulence and virulence factors*

Virulence was once described as the "relative capacity of a microorganism to cause damage in a host" (Casadevall, et al., 1999; Madigan, et al., 2002) and relates to the complex interaction of host and pathogen. Virulence is implemented by so-called virulence factors (VF) which increase the chance to cause damage. VFs often are encoded on mobile elements, prophages, plasmids or genomic regions showing indications of horizontal gene transfer.

A large number of VFs which affect growth in and interaction with the host have been described in *S. pneumoniae* (Mitchell, et al., 2010). Detailed comparison of genomes from related commensal species *S. mitis* and *S. oralis* with the pathogen *S. pneumoniae* revealed that only a few components are preferentially associated with the *S. pneumoniae* (Kilian, et al., 2008; Johnston, et al., 2010; Madhour, et al., 2011; Kilian, et al., 2014; Kilian, et al., 2019; Denapaite, et al., 2010). A few VF especially important for *S. pneumoniae* are described below.

A capsule is also present in other streptococcal species, and the highly variable pneumococcal polysaccharide capsule, the outermost layer of the cell envelope (Dochez, et al., 1917; Heidelberger, et al., 1923), is a crucial, if not the most important, virulence factor (Burnside, et al., 2010; AlonsoDeVelasco, et al., 1995; Mitchell, et al., 2010; Denapaite, et al., 2010). The capsule enables the cell to evade phagocytosis (Roy, et al., 2014; Jonsson, et al., 1985; Johnston, et al., 2013; Avery, et al., 1931; Mitchell, et al., 2010). In *S. pneumoniae*, more than 90 different capsule types are known (Hoskins, et al., 2001; Johnston, et al., 2013; Bentley, et al., 2006; Park, et al., 2007). Most of the gene clusters involved in capsule biosynthesis are located between the genes encoding DexB and AliA (Tettelin, et al., 2015).

The haemolysin pneumolysin (Ply) is present in almost all pneumococci (Kanclerski, et al., 1987; Benton, et al., 1997; Price, et al., 2009; Mitchell, et al., 2010) but can be found rarely in closely related species like *S. pseudopneumoniae* (Kilian, et al., 2019) and some other Gram positive bacteria (Czajkowsky, et al., 2004). Ply has several independent functions: complement activation, stimulation of apoptosis, formation of pores in host cells as cholesterol-dependent cytolysin (Price, et al., 2009; Mitchell, et al., 2010). The Ply gene is often located near a genomic island encoding the major autolysin LytA (Kilian, et al., 2008; Denapaite, et al., 2010), which is not associated with release of Ply (Balachandran, et al., 2001) and can be found in rare cases also in other species (Kilian, et al., 2008).

There are several cell-wall anchored proteins, which are crucial for pneumococcal virulence and can be divided in three classes: choline-binding proteins (CBPs), LPXTG-motif proteins and lipoproteins. CBPs contain repeat domains, which bind to choline of the cell wall (Mitchell, et al., 2010). Choline is located in teichoic acids (TA), which are bound to the peptidoglycan (PG) of the cell wall (wall teichoic acids, WTA) or to the cell membrane (lipo-teichoic acids; LTA) (Bean, et al., 1977; Swoboda, et al., 2010; Fischer, 1997). LPXTG-motif proteins contain a LPXTG-motif, which is, if located near the C-terminus of the protein, covalently linked to the peptidoglycan (Mitchell, et al., 2010).

The first characterized one is the choline-binding protein PspA (pneumococcal surface protein A), which is present in almost all pneumococci, but highly variable in sequence (Hollingshead, et al., 2006). It is able to protect the cell from host immune system by two mechanisms. By blocking adhesion of host complement factors to the cell surface it inhibits removal by opsonophagocytosis (Tu, et al., 1999). Another feature associated with PspA is evasion of the binding to apo-lactoferrin, which can be found in host mucosa and leads to destruction of pneumococci, by sequence variation (Hammerschmidt, et al., 1999; Shaper, et al., 2004). Highly homologous to PspA is the choline-binding protein PspC, which is also referred to as CbpA or SpsA (*S. pneumoniae* secretory immunoglobulin A-binding protein) (Brooks-Walter, et al., 1999; Hammerschmidt, et al., 1997; Rosenow, et al., 1997). This protein is present in about 75% of all pneumococci (Brooks-Walter, et al., 1999). Like PspA, it is able to protect the cell from phagocytosis by detaining the adhesion of complement factors (Li, et al., 2007). An allelic variant of PspC is Hic (factor H-binding inhibitor of complement) (Iannelli, et al., 2002). PspC also facilitates invasion of the mucosa (Zhang, et al., 2000), cerebrospinal fluid (Orihuela, et al., 2004) and adhesion to the vascular endothelium of the blood-brain barrier (Ring, et al., 1998) of human. The pneumococcal choline-binding protein A (PcpA) facilitates adhesion to nasopharyngeal and lung epithelial cells and can be found in nearly all virulent pneumococci (Khan, et al., 2012). The lipoprotein PsaA (pneumococcal surface antigen) is involved in manganese transport (Dintilhac, et al., 1997). Due to manganese and the manganese-dependent superoxide dismutase SodA it contributes mainly to the resistance to oxidative stress (Ogunniyi, et al., 2010). It is highly conserved among main virulent pneumococcal serotypes (Sampson, et al., 1997). Unlike PspA and PspC, it is not likely to elicit opsonic antibodies and thus avoids opsonophagocytosis. This is because of PsaA is anchored at the

outer cell membrane and exposure to antibodies depends on the thickness of cell wall and capsule (Ogunniyi, et al., 2002). The effect of alterations of PsaA is thus quite moderate. Furthermore, the proteins PiuA and PiaA (pneumococcal iron uptake/acquisition) of different uptake systems are involved in virulence and bacterial growth (Brown, et al., 2001). They are conserved and present in all pneumococci (Brown, et al., 2001). Finally, the hyaluronidase (hyaluronate lyase) HlyA which can rarely be found in other species (Madhour, et al., 2011; Kilian, et al., 2019) is present in almost all pneumococci (Paton, et al., 1997). HlyA depolymerizes hyaluronic acid, which is an important component of connective tissue and the extracellular matrix of the host and contributes to colonization (Starr, et al., 2006; Jedrzejas, 2001). It is anchored in the cell wall but can also be released into surrounding host tissue. Many more genes and proteins associated with virulence in have been described in *S. pneumoniae*, but an increasing number of them can also be found in related commensal species (Kilian, et al., 2019). They express important functions that decrease their survival rate in mouse models.

Presence of most pneumococcal VF in commensal species seems to indicate their necessity for colonization and interaction with host tissue.

1.1.1.4 *Penicillin resistance*

One main concern in many bacterial species is the evolution and spread of resistance against many types of antibiotics (Croucher, et al., 2011; Kilian, et al., 2014; Munita, et al., 2016). Antibiotic resistance anciently arose from interaction with the environment and thus many bacteria inherently carry resistances (Munita, et al., 2016). The acquisition and thus spread of resistances by formerly susceptible bacteria is a severe problem in the infectious disease field (Munita, et al., 2016). In streptococci, especially pneumococci, genes involved in the resistances to many antibiotics and their spread throughout the pneumococcal population are well investigated (Garriss, et al., 2019; Schroeder, et al., 2016; Hakenbeck, et al., 1999; Andam, et al., 2015; Reinert, 2009). Acquisition of resistance frequently involves transposable elements, which can carry several resistance determinants. A well-known example are Tn916-family transposons, conferring resistance against macrolides and tetracycline (Schroeder, et al., 2016; Roberts, et al., 2011). Penicillin resistance, on the other hand, is an example where horizontal gene transfer mediated by genetic transformation leads to rapid spread of this new phenotype within the population of *S. pneumoniae* worldwide (Andam, et al., 2015; Reinert, 2009). Resistance is due to a variety of mechanisms: destruction of the antibiotic by hydrolysis or modification, modification of target molecules, or efflux mechanisms. Penicillin resistance in *S. pneumoniae* is achieved by modifications of penicillin binding proteins (PBPs), which are crucial for assembly of the peptidoglycan layer (Hakenbeck R, 1998; Scheffers, et al., 2005; Hakenbeck, et al., 2012), or other components also demonstrated in the current work. *S. pneumoniae* contains six PBPs, which are inhibited by beta-lactam antibiotics, which act as substrate analogue, by forming a covalent complex to the active serine. Thereby, the enzymatic function of PBPs, the transpeptidation of muropeptides of the cell wall, is inhibited, resulting in a less crosslinked cell wall (Hakenbeck R, 1998; Fani, et al., 2014; Munita, et al., 2016). Modification of PBPs is a perfect example of the evolutionary power of horizontal gene transfer followed by recombination events, leading to mosaic gene structure. The introduction of point mutations can alter the affinity to the antibiotics and leading to changes of the resistance profile (Chambers, 1999; Hakenbeck, et al., 2012; Hakenbeck, et al., 1999).

1.1.2 Commensal close relatives of *Streptococcus pneumoniae*

In contrast to its closest relatives *S. pseudopneumoniae*, *S. mitis* and the more distant related *S. oralis*, *S. pneumoniae* is associated with a variety of diseases. Thus, they represent interesting species to investigate the evolution of *S. pneumoniae* and its pathogenic potential. Only in rare cases, *S. mitis* has been identified as the cause of severe diseases mainly endocarditis (Byrd, et al., 2017). The first available complete genome sequence of *S. mitis* was of the high-level penicillin- and multiple antibiotic-resistant strain *S. mitis* B6 (Denapaite, et al., 2010). With 2.15 million nucleotides (nt) (Denapaite, et al., 2010) this sequence is noticeably larger than other *S. mitis* genomes with an average size of 1.8 million nucleotides, which indicates successful incorporation of additional DNA (Kilian, et al., 2008). This genome was analysed to clarify the relationship of *S. mitis* and *S. pneumoniae*. Although *S. mitis* is believed to be naturally competent due to sequence fragments that originate apparently from several sources and the presence of genes necessary for competence and transformation (Salvadori, et al., 2019), the strain B6 shows only low transformation efficiency in laboratory despite the presence of necessary genes (Denapaite, et al., 2010). It contains most of the virulence factors described in *S. pneumoniae*, except some surface proteins, the pneumolysin and the capsule cluster (Denapaite, et al., 2010). Comparison of *S. mitis* B6 with

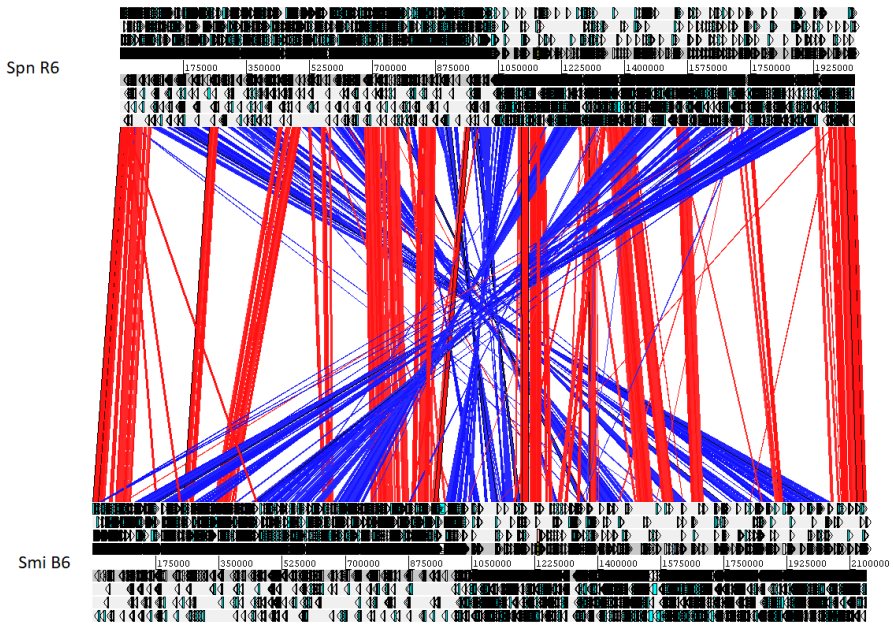


Figure 1.1: Alignment of genomes of S. pneumoniae R6 and S. mitis B6

The visualization with Artemis comparison tool (ACT) shows the alignment of *Streptococcus pneumoniae* R6 (**top**) and *Streptococcus mitis* B6 (**bottom**). Red bars indicate regions with same orientation, blue bars regions with inverted orientation. This arrangement is termed X-alignment.

S. pneumoniae R6 shows an interesting genomic arrangement. Several homologous sequence regions are located at an inverted position in respect to the origin of replication, an arrangement known as X-alignment which has been observed in other organisms as well (Eisen, et al., 2000; Nakagawa, et al., 2003; Denapaite, et al., 2010) (Figure 1.1). Although biofilm formation (Cowley, et al., 2018) and presence of phages (Nakagawa, et al., 2003) are associated with these large-scale rearrangement events, the origin of this phenomenon is still unclear.

S. oralis forms a well separated group distinct from *S. pneumoniae* and *S. mitis* (Reichmann, et al., 2011). The first finished genome sequence is that of the strain *S. oralis* Uo5 , which is, like *S. mitis B6*, high-level penicillin and multiple antibiotic resistant (Reichmann, et al., 2011). *S. oralis* Uo5 was isolated in the 1980s in Hungary (Reichmann, et al., 1997) and is transformable under laboratory conditions (Reichmann, et al., 2011). Similar to *S. mitis* B6, most pneumococcal virulence factors also are present in *S. oralis* Uo5 and when compared to *S. pneumoniae* R6, a noticeable X-alignment can be observed (Reichmann, et al., 2011). An X-alignment is also observed in comparison to *S. mitis* B6, but it is not as distinct when compared to *S. pneumoniae* R6. Alignments with other members of these species are not available since complete genomes are required for such an analysis. Thus it is not clear if this phenomenon is characteric for these species.

The species *S. pseudopneumoniae* was only recently described as a close relative of *S. pneumoniae* and one complete genome sequence of this species is now available: *S. pseudopneumoniae* IS7493 (accession number NC_015875) (Shahinas, et al., 2011). This species can be distinguished clearly by genetic and phenotypic properties like bile solubility, optochin resistance and absence of a capsule from its closest relatives *S. mitis* (NCTC12261) and *S. pneumoniae* (R6). Several recombination events are apparent, which introduced some virulence factors and genes for antibiotic tolerance and resistance as well as surface proteins necessary for host-interaction compared to *S. mitis*, but the absence of crucial virulence factors like the pneumococcal capsule, the choline-binding proteins PcpA, PspA and PcpC, the bacteriocin-like peptide cluster (Blp) and the pneumococcal iron acquisition operon (*piaABCD*) distinguishes this species from *S. pneumoniae*. However, since the species *S. pseudopneumoniae* has gained a certain pathogenicity potential and was already described to cause severe diseases but is genetically located near the commensal members of the Mitis

group streptococci, it seemed to be a good example of an organism, which pursues the thin line between pathogen or commensal (Shahinas, et al., 2011; Shahinas, et al., 2013). A more recent study (Kilian, et al., 2019) revealed that *S. pseudopneumoniae* genomes can be very well classified, but the clinical importance of this species still remains unclear and needs further investigation.

Recombination events within and between streptococcal species result in a large accessory genome and thus in a high variation of genome sequences and gene contents. Thus, different species contain the same genes independent on the expressed virulence and pathogenicity. The difference of virulence and pathogenicity seems to occur from the combination of several alleles.

1.2 Sequencing, assembly and annotation

1.2.1 Sequencing technologies

During the last four decades, several sequencing technologies were developed to obtain the sequence information of bacterial genomes. These DNA sequencing technologies provide data for a broad field of analyses like evolution and comparative analysis, forensics, health care (diagnostics, antibiotic resistance, etc.)(for example (Kilian, et al., 2008; Ranjan, et al., 2017; Wallace, et al., 2016; Cohen, et al., 2015; Arigmón, et al., 2014; Alvarez-Cubero, et al., 2017; Cao, et al., 2017)). The ongoing improvements of these technologies as well as reduction of costs and time requirements facilitate the generation of a vast increasing number of genomic data.

There are several technologies to obtain DNA sequences, starting 1977 with the chain termination (or dideoxy-) method of Sanger and the Maxam-Gilbert-method (Maxam, et al., 1977; Sanger, 1977) of base specific cleavage, which introduced the first of the current three generations of sequencing technologies (Land, et al., 2015).

The dideoxy-method, emerging from the inaccurate “plus and minus”-method, uses a low concentration of labelled dideoxy nucleotide triphosphates (ddNTPs) besides “normal” deoxy nucleotide triphosphates (dNTPs) during DNA synthesis followed by electrophoresis of the generated DNA molecules (Sanger, 1977). During amplification of the sample, DNA-polymerase I incorporates dNTPs into the replicated strand to elongate the copy. Additionally, a small volume of ddNTPs is added during amplification. Due to the lack of the 3'-hydroxyl group of ddNTPs, the DNA-polymerase is not able to elongate further, and the elongation is terminated. Since ddNTPs are incorporated randomly, DNA-fragments of several lengths are generated for each of the four types of nucleotide. Subsequent polyacrylamide gel-electrophoresis of the four fragment sets reveals the order of incorporated nucleotides and thus the sequence.

In contrast to the enzymatic chain-termination method, which determines a sequence during amplification, the Maxam-Gilbert-method determines the sequence by fragmentation of DNA. Originally, sample DNA is thus labelled radioactively at one end and then modified and cleaved at four lanes with base specific (A, G, C, C+T) reagents (Maxam, et al., 1977). Length determination with polyacrylamide gel-electrophoresis then reveals the base sequence.

Although the automated Sanger method was the main sequencing technology for more than twenty years, increased demand as well as high cost and duration of first-generation technologies led to further improvements and finally to development of parallel operating second-generation technologies, also called next-generation technologies (NGS) (Metzker, 2005; Metzker, 2010; Lu, et al., 2016; Miller, et al., 2010). The most popular methods are the pyrosequencing method of 454 Life Sciences/Roche (Margulies, et al., 2005; Schatz, et al., 2010), the bridge synthesis method of Illumina/Solexa (Hillier, et al., 2008; Liu, et al., 2012; Bentley, et al., 2008) and the two-base method of ABI/SOLiD (sequencing by oligo ligation detection) (Liu, et al., 2012; Valouev, et al., 2008) but also sequencing by hybridization (DNA chip/microarray) (Liu, et al., 2012; Lipshutz, et al., 1995; Lipshutz, et al., 1995) and ion-torrent semi-conductor sequencing (Liu, et al., 2012; Rothberg, et al., 2011). Since the 454 and the Illumina technology delivered data used in the current work, they are described more detailed.

The 454-method was the first commercially and large-scale used technology of the second generation, but although still used it is about to be more and more displaced by Illumina-technology. The sample DNA is amplified by emulsion PCR (emPCR). This kind of amplification uses an emulsion of sample DNA, beads, primer, dNTPs and polymerase to multiply the single stranded sample DNA, which then is bound to the beads. Afterwards, these beads are placed onto a picotiter-plate with one bead per well. Smaller beads loaded with immobilized ATP-sulphurylase and luciferase are added to each well. Then, in a circular manner, all four dNTPs are added sequentially. DNA-Polymerase incorporates the nucleotides and releases pyrophosphates, which are converted to ATP (adenosine-triphosphate), which serves as substrate for luciferase, which converts to oxyluciferin under light emission. After each nucleotide flow, the emitted light signal is detected, and the number of incorporated nucleotides is defined by the signal intensity. After signal detection, the next nucleotide flow starts (Metzker, 2010; Margulies, et al., 2005).

Although the 454 technique generates reads with a higher average length, the results are afflicted with errors arising from rounding of signal intensity values of homopolymer stretches (HPN) as described in chapter 0.

Different to the 454-technology, the Illumina-technology incorporates only one nucleotide per flow and another amplification procedure (Bentley, et al., 2008). The 'bridge amplification' takes place on a solid surface where oligonucleotides are bound. The same (and complementary) oligonucleotides also are ligated to the end of double-stranded sample DNA-fragments. The DNA is denatured, and the oligonucleotide of the single-stranded sample DNA then binds to the complementary and immobilized oligonucleotide. The complementary strand of the sample DNA then is synthetized and extends the immobilized oligonucleotide, after which the original strand is removed. The free end of the strand binds to immobilized oligonucleotides nearby, forming a 'bridge' (Figure 1.2). Denaturation of the DNA and subsequent repeat of this process leads to the amplification of the sample DNA and formation of clusters on the solid surface. For sequencing, added nucleotides are labelled at the 3'-end with fluorophores for detection (different for each kind of nucleotide) and prevention of further nucleotide-incorporation (termination). Thus, all four kinds of nucleotide can be added at once, but only one is incorporated by polymerase and can be detected by its specific label. After detection (by laser) the fluorophore is removed and further nucleotide-incorporation in the next flow is possible.

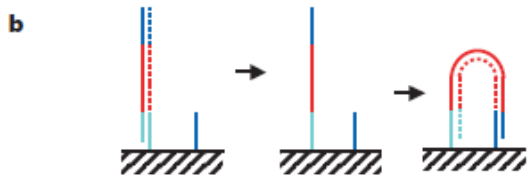


Figure 1.2: Schematic representation of bridge amplification

Sample DNA binds to immobilized oligonucleotide. Subsequent generation of complementary strand and binding to another oligonucleotide nearby leads to 'bridge'. Separation of the now double-stranded DNA und repeating of this process amplifies the DNA and forms clusters of copies of the same DNA sample.

On the one hand, the popular second generation procedures for sequencing genome-size DNA (454, Illumina, SOLiD (454 read length improved)) produce higher coverages than the first-generation technologies, but also shorter reads with an higher error rate (Hutchison, 2007; Liu, et al., 2012; Fullwood, et al., 2009). While the higher error rate might be compensated by the higher coverage, the brevity of the reads often is the cause of assembly problems at repetitive sequence regions, which cannot be spanned by short reads. To reduce this disadvantage of these methods, it is possible to combine them with paired-end-sequencing (Illumina generally works with paired reads). At this technique, pieces of DNA with certain (and known) length are circularized and fragmented (Figure 1.3). The fragments containing the ligated ends of each sequence piece are sequenced and provide a pair of sequence reads, whose distance is known. This information can be used in an assembly to close gaps, which cannot be spanned by short single reads (see chapter 1.2.2).

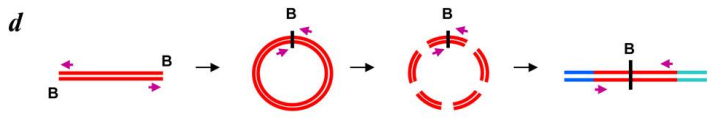


Figure 1.3: Schematic representation of generation of paired reads

Biotinylated adaptors are added to both ends of double stranded DNA to facilitate binding of the two ends to each other. Nebulization and subsequent washing retrieves fragments containing the biotin label and thus the sequence fragments, which are located at the ends of the original sequence.

Source: (Bentley, et al., 2008), Figure 1

The latest (third) generation of sequencing technologies (TGS) handles larger read lengths and facilitates spanning over repeats and is represented by Pacific Biosystems (PacBio) and Oxford nanopore technology (ONT) (Ashton, et al., 2015; Cao, et al., 2017; Eid, et al., 2009; Mayjonade, et al., 2016; Chen, et al., 2015; Laver, et al., 2015). These technologies work with average read lengths about 3k, but can reach lengths of several ten or hundred thousand nucleotides per read (Lu, et al., 2016; Chen, et al., 2015). One recent example of third generation sequencing is the amplification-free retrieval of the M13 virus genome with a size of about 3.700 nt using single molecule sequencing technology (Zhao, et al., 2017).

As example, the ONT is explained here. The library preparation of the ONT differs massively from the NGS preparations. At first, sample DNA is sheared and repaired if necessary (Lu, et al., 2016). Then adapters are added to the two ends of the molecule: A leader adapter at the 5'-end and a hairpin adapter at the 3'-end (connects both strands if dsDNA is used) (see Figure 1.4). Starting at the leader adaptor, dsDNA is unzipped, and the forward strand is led through

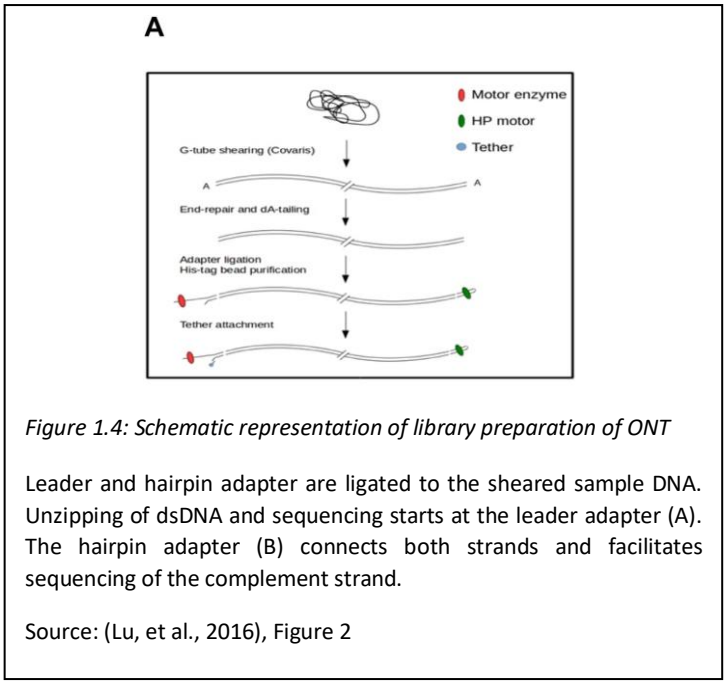


Figure 1.4: Schematic representation of library preparation of ONT

Leader and hairpin adapter are ligated to the sheared sample DNA. Unzipping of dsDNA and sequencing starts at the leader adapter (A). The hairpin adapter (B) connects both strands and facilitates sequencing of the complement strand.

Source: (Lu, et al., 2016), Figure 2

the nanopore. When the hairpin adapter is reached and if dsDNA is used as sample, the complement strand is also led to the nanopore. Nanopores are located at a membrane (512 pores per membrane) where a voltage is applied. A DNA-molecule, which passes a nanopore, changes the current, which can be measured (several thousand times per second). This measurement leads to sequences of "events" (changes of the ion current), whereof 5- or 6-mers are calculated with a Hidden Markov Model (see chapter 1.2.2) to generate a path representing reads of the sample DNA. These reads initially can contain base errors of 25 – 35% at ssDNA (1D-reads) and 12 – 20% at dsDNA (2D-reads). For comparison, the error rate of the PacBio technology is about 10 – 15%. Subsequent error correction is able to reduce this error rate to about 0.5% as demonstrated at the genome of *E. coli* K-12 MG1655.Further developments and improvement of this sequencing technology try to reduce the initial error rate.

Beginning with two sequenced bacterial genomes in 1995 (Haemophilus influenzae (Fleischmann, et al., 1995) and Mycoplasma genitalium (Fraser, et al., 1995)), the first *Streptococcus pyogenes* (M1) (Ferretti, et al., 2001) and *Streptococcus pneumoniae* (TIGR4 and R6) (Tettelin, et al., 2001; Hoskins, et al., 2001) genomes were published in 2001. Improvement of sequencing technologies and reduction of costs and sequencing duration during the last two decades led to tens of thousands of bacterial genomes today, assembled from a vast number of sequence data (Land, et al., 2015). The *S. pneumoniae* data base alone (https://www.ncbi.nlm.nih.gov/genome/genomes/176) lists 8.514 of genomes to date (November 2019).

1.2.2 Genome assembly

The sequencing techniques used for the current work belong to the next generation sequencing (NGS) and thus, the assembly process of NGS generated data will be described here.

Sequencing produces sequence reads representing the nucleotide sequence and further information like quality or pair information. To obtain a genome sequence, these fragments have to be assembled, which is possible, when the sequence information of all sequence reads covers the size of the real sequence and if there is an overlap of the reads. However, if sequence reads are shorter than repetitive sequence regions, which are present in almost all bacterial genomes, the assembly of this region is problematic as described below and has to be solved otherwise. The task of an assembly is the grouping of sequence reads into contigs and scaffolds (also called supercontigs or metacontigs), where contigs represent multiple alignments of reads and their consensus sequences and scaffolds represent these contigs in defined orientation and order as well as the size of gaps between contigs (Miller, et al., 2010).

The assembly of NGS data is generally performed by NGS assembler software. This kind of software is based on so called graphs and can be divided into three groups using several forms of graphs: De Bruijn graph (DBG), overlap/layout/consensus (OLC) and greedy graph. Common to assemblers using these types of graph is a pre-processing of reads to reduce possible errors, a simplification during or after graph generation, including usage of information from outside the graph, and the generation of consensus sequences for contigs and scaffolds (Miller, et al., 2010).

A graph is a set of nodes and edges and can be used as abstraction for sequence data processing. A collection of edges visiting nodes in a certain order is called path. The complexity of a graph is determined by the size and repeat structure of the sequenced genome (Miller, et al., 2010; Nederbragt, 2010). According to the underlying sequence, the graph may contain so called spurs, which are diverging dead-end branches, for example induced by sequencing errors at read end or zero coverage. Bubbles are diverged branches, which converge back to the “main” branch, induced by mid-contig sequencing errors and polymorphism. Repeats might induce a frayed rope pattern when a converged branch is diverging again. They also

might induce cycles, paths converging on themselves. Such divergences and convergences enlarge the complexity of a graph (see Figure 1.5) (Miller, et al., 2010).

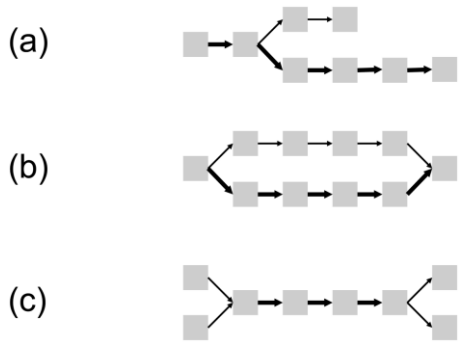


Figure 1.5: Schematic representation of graph complexity

Nodes of a graph might be connected not only in a straight sequential manner, but also in more complex forms like (a) "spurs", (b) "bubbles", (c) "frayed rope" or other.

Source: Miller *et al.* 2010, Figure 3.

OLC graphs needs pre-calculation of all-against-all pairwise alignments of sequence reads. The graph itself represents the reads (nodes) and their overlaps (edges). Paths through the graph form potential contigs. Usually, the workflow of assemblers using OLC graphs consists of three main steps. During the first step overlaps of all reads are estimated. After this, a graph is constructed and adjusted. The third step is a multiple alignment and the generation of a consensus sequence per path and thus contig. An assembler using OLC graphs for example is *Newbler* (Margulies, et al., 2005; Miller, et al., 2010).

Originally developed for representation of string sequences, the DBG represents all possible fixed-length substrings at nodes and identical suffixes and prefixes of nodes at edges. A special and popular form of DGB is the K-mer graph (K-mer = substring of length *K*), using a fixed length of pre- and suffixes. The nodes of a K-mer graph during assembling WGS data represent sequence reads, while edges represent identical pre- and suffixes of these reads. Due to these identical alignments, K-mer graphs are more sensitive to repeats and sequencing errors than OLC graphs. An assembler using DBG/K-mer graphs is for example *Velvet* (Zerbino, et al., 2008; Miller, et al., 2010).

Assemblers using greedy graphs use a very stringent form of OLC or DBG graph. A read is only joined with a read with the highest score. This is repeated while possible. An assembler using the greedy algorithm is for example *SSAKE* (Warren, et al., 2007; Miller, et al., 2010).

The certainty of assemblies can be improved by so called paired ends (also called mate pairs), which also aid in gap closure and linkage of contigs to scaffolds. At least from 1981 on (Hong, 1981), paired ends are used in variations for several applications (Fullwood, et al., 2009; Hong, 1981; Collins, et al., 1984; Fleischmann, et al., 1995; Bentley, et al., 2008). Common to all is the core principle of circularizing DNA of a certain length, where the latter paired sequence fragments are linked and sequenced together. During assembly, such read pairs can be used for estimation of the average distance between paired sequences, when both fragments are located within the same contig (Miller, et al., 2010; Nederbragt, 2010). This information can be used to estimate the gap size between not overlapping contigs, where each contig contains a fragment of at least one pair (Miller, et al., 2010; Nederbragt, 2010). These gaps are filled up with the indicator for ambiguous nucleotides ('N') (Miller, et al., 2010; Nederbragt, 2010; Fullwood, et al., 2009).

An alternative to a de novo assembly is a mapping of the reads using a reference sequence. This is useful to detect single nucleotide polymorphisms (SNPs, substitutions) and deletion and insertion polymorphisms (DIP, indel) between closely related organisms (Miller, et al., 2010).

The assembly procedure in detail is dependent on the used assembler. Most of the assemblies used for the current work were performed with *gsAssembler* (*Newbler*) (Margulies, et al., 2005) version 2.6 and thus some properties of the procedure are described here. Initial to the assembly, *Newbler* generates so-called seeds from each read (Nederbragt, 2010). These seeds are sequence fragments of a certain length and position within a read and thus, each read contains a defined set of seeds. Differ two reads in their sequence but not in their seed-set due to differences in the sequence between seeds, then seeds can be extended until the two sets differ. This happens dynamically and automatically during joining reads at similar or equal ends (dependent on program parameters). Reads are removed from assembly, if their ends do not overlap with other reads (singletons) (454 Life Sciences Corp, 2009). Furthermore, Newbler removes sequence repeats from assembly, when more than 70% of the seeds of one read have an identity of at least 70% to the seeds of another read. Apart from that, partial repetitive reads are used in the assembly. So-called outlier, problematic reads e.g. due to chimeric sequence, are also removed from assembly as well as too short reads. The result of removal and joining of reads are so-called contigs (from 'contiguous'), sequence fragments composed of several overlapping reads (compare description of graphs above). A problem

occurs, when overlapping read fragments continue in different sequence contents, e.g. because the single sequences might be located at distant positions within the genome. In this case the consensus sequence of the reads is split, and the fragments are assigned to several contigs (Figure 1.6). Especially repetitive sequences or such with a high number of copies within the genome like insertion sequences (IS) or RNA are often cause of segregation of contigs and reads. After the assembly has finished, the resulting contigs are written into output files.

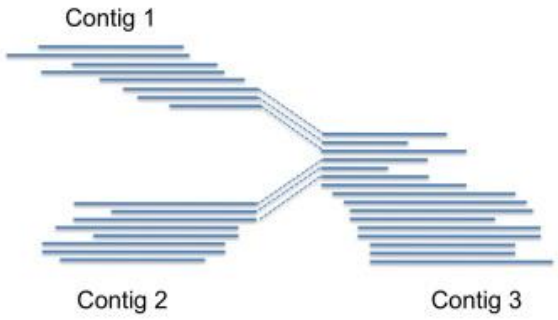


Figure 1.6: Scheme of reads spanning several contigs

Partially overlapping reads lead to a branching of the graph and the consensus sequence is thus split into fragments which are assigned to several contigs.

Source: Nederbragt, 2010

1.2.3 Sequence errors

Based on the used sequencing technology as well as the sequenced object, several errors might occur during assembly (and preceding sequencing).

A major problem of assemblers are repeat regions, especially at the processing of short reads, resulting in ambiguous assignment of reads. Some repeats might be bridged by so called spanners, reads longer than the repeat and containing unique sequences at each end. Furthermore, paired reads might contribute to the solution of this problem, when one member of a pair is located in unique sequence outside the repeat. If the repeats are not exact, a more stringent alignment might also reduce the problem. In general, shorter reads are not appropriate to solve repeat regions. The problem is worsened by sequencing errors. Increased error tolerance of the assembler software would increase the rate of correctly assembled reads, but also the rate of false positive alignments, especially at non-exact repeats. While sequencing technologies do not provide error-free data, assemblers cannot work stringent and a certain uncertainty regarding repeats will continue. Repeats, polymorphisms and sequencing errors may lead to increased graph complexity during assembly (Miller, et al., 2010). For the first complete genomes, gaps were filled by manually sequencing using primers that match the ends of contigs (see (Hoskins, et al., 2001; Reichmann, et al., 2011; Denapaite, et al., 2010)).

Detection of the signal intensity during 454 sequencing causes inaccuracies due to the underlying technology (Margulies, et al., 2005; Brockman, et al., 2008; Luo, et al., 2012; Ronaghi, 2001). Read errors occur by over- or undercalls rather than miscalls during flowgram calculation (Huse, et al., 2007). Over- or undercalls emerge, when the calculated value differs by at least 0.5 units from the real amount of the affected type of nucleotide and rounding errors are generated. The rounding errors cumulate by increasing stretch length of one type of nucleotide. Consequently, the sequences of the generated contigs from different genomes may contain homopolymer stretches of different lengths in homologous regions, which may result in artificially disrupted genes. For sequence confirmation, these stretches have to be determined by another sequencing method. Some assemblers like Newbler are able to reduce the amount of such under- and overcalls by calculating more precisely the length of a homopolymer stretch at increasing coverage (Miller, et al., 2010). At Illumina/Solexa sequencing, error rates in homopolymer stretches are not increased compared to other

sequence regions albeit polyA- and polyT-stretches contribute to an increased overall error rate by protracted (carry forward) errors, as well as repeat regions as mentioned above and GC-content of the sequenced genome (Dohm, et al., 2008).

Using Illumina/Solexa sequencing technology, detection errors also may occur. Different from 454 technology, where only one type of nucleotide is added per time, all four bases are added per time for sequence generation and detection at this technology. The nucleotides G and T are detected with the same laser, distinguished only by signal strength, as well as the nucleotides A and C. Thus, it is not surprising, that substitution of G by T and A by C seem to be the most frequent substitution errors (26 - 43% (Dohm, et al., 2008)). It was found, that errors are preferentially located at positions after a G-rich sequence region, which indicates problems of incorrect flushing between the sequence elongation steps and thus incomplete de-protection and removal of fluorophore from the formerly added nucleotide (Dohm, et al., 2008). Protected nucleotides inhibit the incorporation of the subsequent nucleotide and lead to erroneous sequence data (Dohm, et al., 2008). In addition, one or more Gs prior to a SNP hint at a possibly wrong base call (Dohm, et al., 2008).

The overall error rate differs depending on the used sequencing technology. So, the error rate of 454 generated sequences concerning indels is approximately 0.31% (50% of which are in homopolymer stretches) (Huse, et al., 2007) and in sequences generated by Illumina/Solexa technology less than 0.01% (thereof about 25% in homopolymer stretches with at minimal length of four) (Dohm, et al., 2008).

As ambiguous bases generally indicate erroneous reads, removal might drop the overall error rate and improve assemblies and mappings (Huse, et al., 2007).

A difference between the 454 and Illumina/Solexa sequencing technologies is the different meaning of the per base quality score provided with each of these methods. While the 454 score indicates the probability of correct homopolymer length, the Illumina/Solexa score indicates the probability of correct base call. Despite the different meaning of quality scores, higher values indicate fewer error rates (Huse, et al., 2007; Dohm, et al., 2008).

1.2.4 Recognition and annotation of genomic features

There are several methods and programs to identify and describe structures within a genome sequence, i.e. to annotate a given sequence. These structures can be protein coding sequences (CDS), RNA genes, repeat regions, phages or other genomic features. Examples for annotation software are the *PGAP* (NCBI prokaryotic genome annotation pipeline) (Tatusova, et al., 2016) or the *RAST* (rapid annotation using subsystem technology) pipeline (Aziz, et al., 2008). Both pipelines employ partially different but also similar methods and algorithms to determine genomic structures. These methods are described here representatively.

For prediction of protein coding genes, both pipelines use programs (e.g. *Glimmer* (Salzberg, et al., 1998; Delcher, et al., 1999) and *GeneMarkS* (Besemer, et al., 2001)), which are based on so called Markov models (MM). In general, MM are sequences of random variables, whose occurrence probability is only determined by a certain number of preceding variables (Delcher, et al., 1999). Applicated for DNA sequences, this means that the probability of one nucleotide is determined by its predecessors (Delcher, et al., 1999). These preceding nucleotides are called the context of this nucleotide (Delcher, et al., 1999). There are several variants of MM, where 5^{th}-order MM (five preceding nucleotides; fixed order MM) have turned out to be effective for the prediction of bacterial genes (Delcher, et al., 1999; Audic, et al., 1998; Borodovsky, et al., 1995). Disadvantage of these fixed-order MM is that they are only reliable if enough training data are available (Salzberg, et al., 1998; Delcher, et al., 1999). To solve this problem, interpolated MM (IMM) use different context lengths up to 8^{th} order only with enough training data. The interpolation of a linear combination of probabilities calculated from different context lengths and context weight, which depends on the oligomers occurrence frequencies, result in more accurate gene predictions than with fixed-order MM (Salzberg, et al., 1998; Delcher, et al., 1999). In addition, *Glimmer* and *GeneMarkS* produce three IMM (3-periodic MM) per strand to cover all three frames (Salzberg, et al., 1998; Delcher, et al., 1999; Besemer, et al., 2001). To resolve gene overlaps, *Glimmer* compares the scores of the single genes and the overlap region and, if possible, moves the start position of the genes, to decide which of both genes remains in the annotation (Salzberg, et al., 1998; Delcher, et al., 1999). In contrast, *GeneMarkS* employs heuristically scored models (heuristic MM; HMM) using e.g. ribosomal binding sites (RBS) and spacer lengths (sequence between RBS and supposed gene start) to predict correct gene starts (Besemer, et al., 2001). Compared

to each other, *Glimmer* and *GeneMarkS* identify about the same number of genes (Delcher, et al., 1999; Besemer, et al., 2001).

To detect RNA genes, other approaches are applied. For transfer RNA (tRNA), tools like tRNAScan-SE (Lowe, et al., 1997) are used. *tRNAScan-SE* was designed to determine eukaryotic RNA genes but is also usable for other organisms (Lowe, et al., 1997). This program works in three steps. At first, the core *tRNAScan* program (Fichant, et al., 1991) is applied together with an algorithm (Pavesi, et al., 1994), which determines tRNA genes by recognition of two intragenic control sequences, the transcription termination site and the spacer between them (Lowe, et al., 1997; Pavesi, et al., 1994). The covariance search model program *covels* (Eddy, et al., 1994) then tries to validate the predicted tRNA genes with the *Spinzl* database (Steinberg, et al., 1993), which contains sequences of tRNA genes (Lowe, et al., 1997). During the third step, the validated sequences are used by the covariance model global structure alignment program *coves* (Eddy, et al., 1994) to predict secondary structures, where anticodons are tried to be determined (Lowe, et al., 1997). The usage of heuristic data helps to determine pseudogenes (Lowe, et al., 1997).

Besides homology searches in databases (Tatusova, et al., 2016) like *Rfam* (Griffiths-Jones, et al., 2005; Nawrocki, et al., 2015) or usage of private tools (Aziz, et al., 2008), already existing tools like *Infernal* (Nawrocki, et al., 2013), which allows search of RNA in databases and generation of multiple sequence and structural alignments of RNA and is based on hidden Markov models (HMM) and covariance models, are applied to determine putative ribosomal RNA (rRNA) sequences.

The PGAP pipeline identifies core proteins of the pan-genome of a specific prokaryotic clade, which is predetermined or determined by ribosomal markers. After clustering them with *USearch* (Edgar, 2010), the proteins of this set are used as so-called footprints for *GeneMarkS* (Besemer, et al., 2001) to determine proteins in a new genome. Furthermore, RNA genes are determined using *Rfam* (Griffiths-Jones, et al., 2005; Nawrocki, et al., 2015) and *BLASTN* (Altschul, et al., 1990) (see below) and a family of repetitive regions (clustered regularly interspaced palindromic repeats; CRISPRs) with the especially designed CRISPR recognition tool *CRT* (Bland, et al., 2007) and *PILER-CR* (Edgar, 2007). Phages are recognized using a database containing phage and plasmid proteins with *TBLASTN* (Altschul, et al., 1990) and *ProSplign* (Sayers, et al., 2011). Protein genes detected by these previous steps are passed as

footprints or hints to *GeneMarkS* for further determination and confirmation, where also RNA genes are taken into account (Tatusova, et al., 2016).

The *RAST* pipeline uses *Glimmer2* (Delcher, et al., 1999) to identify proteins independent on the correct start codon. These predicted proteins are compared to a set of FIGFams, which are universal in most prokaryotes to determine the closest neighbours of the current genome. A FIGFam is a set of proteins, which share a family function and a decision procedure for adding new proteins. The predicted proteins are searched within the FIGFams of the closest neighbours and the start positions as well as overlaps of the encoding genes are corrected if necessary. Predicted proteins, which cannot be found among this set of FIGFams, are searched against the whole manually curated FIGFam database. tRNA genes are determined by usage of *tRNAScan-SE* (Lowe, et al., 1997), while rRNA genes are determined by an internal tool called *search_for_rnas*. Besides download of the annotated sequence, browsing and comparing this genome with other genomes is provided by *RAST*, as well as download of additional information like subsystem collection, where a subsystem is a set of functional roles, which describe special biological processes or complex structures (Overbeek, et al., 2005; Aziz, et al., 2008).

Another attempt of annotating genomes is the transfer from a reference sequence by extraction of annotated features, which are searched for in the sequence to be annotated. This attempt needs no underlying database and corresponding tools are introduced by *RATT* (rapid annotation transfer tool) (Otto, et al., 2011).

Looking at the described excerpt of annotation pipelines and software, the underlying methods are rather similar, although differing in their final application and further improvements. The choice, which tool to use for annotation, depends on the current needs of the annotating researcher. Genomes used in the current work were annotated using *RATT*, *RAST*, manual transfer from similar genomes or databases (in case of small RNA or finalization). Although the annotation software is still improving and becomes more reliable, none of the methods are free of errors.

1.3 Analysis

Depending on the particular question, there is a variety of possibilities to analyse the retrieved sequence information. Besides statistical analyses like determination of GC-content which can be used to determine the DNA melting temperature (Yakovchuk, et al., 2006) or potential sites of recombination (Lassalle, et al., 2015)), codon usage (relevant to predict translation efficiency) (Chaney, et al., 2015; Hockenberry, et al., 2014) and other (Song, et al., 2014), mostly sequence alignments are used. Alignments are comparisons of character strings, which might represent DNA, RNA or amino acid sequences. During such an alignment of two sequences and based on substitution matrices and gap introduction and extension costs, one

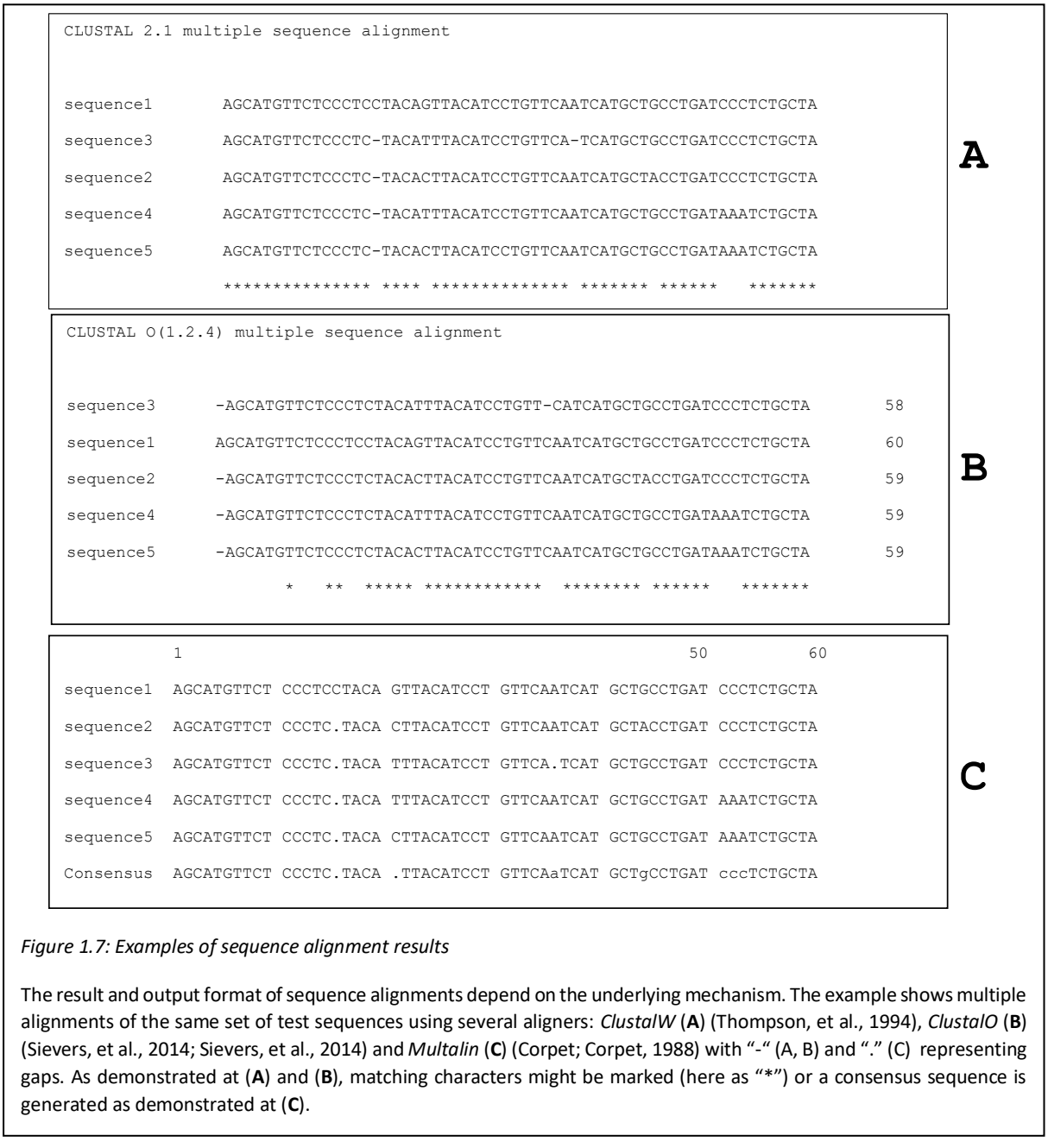


Figure 1.7: Examples of sequence alignment results

The result and output format of sequence alignments depend on the underlying mechanism. The example shows multiple alignments of the same set of test sequences using several aligners: *ClustalW* (**A**) (Thompson, et al., 1994), *ClustalO* (**B**) (Sievers, et al., 2014; Sievers, et al., 2014) and *Multalin* (**C**) (Corpet; Corpet, 1988) with "-" (A, B) and "." (C) representing gaps. As demonstrated at (**A**) and (**B**), matching characters might be marked (here as "*") or a consensus sequence is generated as demonstrated at (**C**).

sequence is tried to be converted into the other sequence (Henikoff, et al., 1992; Miyazawa, et al., 1993; Koshi, et al., 1995; Schneider, et al., 2005; Dayhoff, et al., 1978; Agrawal, et al., 2011; Edgar, 2009; Waterman, et al., 1992). If necessary, gaps can be introduced for alignment purposes. The required steps of this transformation determine the alignment score. The consensus sequence is a representation of the alignment result and describes each compared character position using sequence characters or similarity symbols (match, mismatch, gap, etc.) (Waterman, 1986). Several tools (e.g. *ClustalW* (Thompson, et al., 1994), *Clustal Omega* (*ClustalO*) (Sievers, et al., 2014; Sievers, et al., 2014), *Multalin* (Corpet; Corpet, 1988), *Blast* (Altschul, et al., 1990)) are available for sequence alignments and a variety of possible applications like usage during assembly (see above), searches of sequences or patterns (see below) within single sequences or sequence collections (see https://blast.ncbi.nlm.nih.gov/Blast.cgi). Depending on the underlying algorithm and matrix, the results can differ in content and format (see Figure 1.7). The several alignment algorithms can be broadly separated into local and global alignments. Global alignments calculate comparison and score for the whole sequences and are used for sequences of similar length and assumed high homology (Polyanovski, et al., 2011). At the other hand, local alignments determine the fragments of both sequences with the highest score, which otherwise would get lost in the global score. The most established and basic algorithms are the Needleman-Wunsch algorithm (Needleman, et al., 1970) and the Smith-Waterman algorithm (Smith, et al., 1981).

A multiple sequence alignment is an extended form of the pairwise alignment of two sequences, where alignments are processed in any (e.g. progressive or iterative) manner, to produce reasonable results in improved processing time (Mount, 2001).

A further form of alignment and also used in many other fields of daily life is the search for DNA and protein sequences with patterns. Patterns are sequences providing a certain variability by using wildcards and variable lengths and are used by a variety of tools like *PatMatch* (Yan, et al., 2005) or tools of the *EMBOSS* software suite (Rice, et al., 2000).

Pairwise and multiple alignments facilitate the determination of exchanges, deletions or insertions of single nucleotides (single nucleotide variations, SNVs) referred to a reference sequence, but also of larger sequence regions and make statements possible concerning relationship of several strains or species as well as evolution and horizontal gene transfer.

Analyses are not restricted to the raw sequence, but also applicable to the level of identified or annotated genes and derived proteins. For example, genotyping methods like MLST (multi locus sequence typing) (Enright, et al., 1999) are capable to identify strains initially and assign them to certain taxa depending on the number of sequence deviations (Xu, et al., 2011; Margos, et al., 2018). One question that can be addressed now given the large number of genome data of individual bacterial species is the definition of the core genome (genes present in all strains of a species), the accessory (or dispensable) genome (genes present only in some strains and their putative origin) and the pan genome, representing all genes of (Tettelin, et al., 2005; Guimarães, et al., 2015)). For the definition of the core genome, it is reasonable to tolerate a certain variability of the sequences, i.e. define a minimum similarity (Pearson, 2013).

1.4 Visualisation

There are several tools for visualization of sequences, sequence comparisons/alignments and so on. The ones mainly used in the current work are the tools *Artemis* (Carver, et al., 2012) and the Artemis comparison tool (*ACT*) (Carver, et al., 2005). *Artemis* is capable of displaying single protein and DNA/RNA sequences – also whole genomes – including several features like CDS or genes. Furthermore, it contains a broad palette of functions for analysis, editing and annotation. *ACT* extends the functionality of *Artemis* by displaying two or more sequences and their differences or homologies (*BLAST* table). An example can be seen at Figure 1.1. Both programs run platform-independently due to their implementation in Java and without installation.

1.5 Goals and work objectives

The work presented here concerns detailed genome comparisons of clones and strains of the species *Streptococcus pneumoniae* and with related species in respect of similarities and differences also in regard to structure and spread of virulence and resistance.

The possibility to generate genome sequences with increasing speed and quantity leads to problems regarding the enormous and barely to handle amounts of data. Besides the quantity of data, the underlying sequencing technologies bring along their own problems. Two representatives of such technologies are the 454- and the Illumina-technology. In the present work, genomes were generated by either one of these sequencing methods to investigate changes in the assembled genomes, that are observed between strains belonging to one clone of *S. pneumoniae* (publications I (Rieger, et al., 2017) and II (Rieger, et al., 2017)), that occur in one *S. pneumoniae* strain after transformation with DNA from other species (publication III (Todorova, et al., 2015)) and genomic differences between streptococcal species of the viridans group with emphasis on virulence factors described in *S. pneumoniae* and cell surface components (publications IV (Denpaite, et al., 2016) and V (Tettelin, et al., 2015)). The challenges in such analyses include the definition of SNPs and Indels, and regions of high variability signifying potential sites of recombination with DNA of other species as the result of horizontal gene transfer. Moreover, the identification of altered sites due to sequencing errors which is especially important for SNP definition has been addressed, an issue which is often neglected in many publications.

Sequences described in publications (I) and (IV) were sequenced with 454-technology and Illumina-technology has been applied in publication (II) and (III). Since all sequences were assembled with the same software, the *gsAssembler* of Roche, also known as *Newbler*, which initially was developed to process 454-Data, the question was whether it is possible to compare the used datasets.

Analyses performed after, during and prior to assembly depend on their purposes. Many of them include sequence alignments in a more or less stringent manner. While for example SNV (single nucleotide variations) analysis needs strict settings to not miss single nucleotide mismatches, analyses of core or dispensable genomes need to be more tolerant against differences to a certain degree to bundle genes or proteins with a similar sequence.

Streptococcus pneumoniae, a naturally transformable organism and the main target of the present work, represents a perfect example to study genomic variability within clones and between closely related species to further our understanding on the evolution of antibiotic resistance and the pathogenicity potential associated with this particular bacterium.

In publication I and its unpublished material, SNP retrieval as well as unaligned regions were analysed in detail, and the gene content was examined to investigate whether strains of one *S. pneumoniae* clone ST10523 associated with a cystic fibrosis patient differ from a strain of the same clone that was obtained from a non-cystic fibrosis host. Differences of virulence factors were of special interest due to their contribution to survival within a host. Similar analyses were performed in three strains of a high-level penicillin and multiple antibiotic resistant *S. pneumoniae* clone Hu19A-^{6}, a rare example where one member of a clone was antibiotic sensitive (II). In this case, the focus was on components involved in penicillin resistance in addition to virulence factors. The identification of highly variable regions was important to specify regions introduced via transformation and recombination using the laboratory strain *S. pneumoniae* R6 as recipient (III). The definition of the core genome was important to determine species specific features comparing a wide variety of streptococcal species in publication (IV) and to clarify the speciation of *S. pneumoniae* and closely related species as reviewed in (V).

In all publications emphasis was also placed on penicillin binding proteins (PBPs), a paradigm for mosaic genes as a result of horizontal gene transfer involving different species.

2 Scientific papers

2.1 Long persistence of a novel *Streptococcus pneumoniae* 23F clone in a cystic fibrosis patient

Martin Rieger, Harald Mauch and Regine Hakenbeck. mSphere Jun 2017, 2 (3) e00201-17; DOI: 10.1128/mSphere.00201-17

Summary:

Over a period of 37 months, seven streptococcal isolates were extracted from a cystic fibrosis (CF) patient. All isolates showed intermediate penicillin resistance and belonged to the *S. pneumoniae* serotype 23F clone ST10523. Since *S. pneumoniae* is not known to be a persistent colonizer, the first (D122) and the last (D141) isolate were sequenced to investigate genomic differences, especially in respect to pneumococcal specific virulence factors, that might explain the unusual long persistence of this clone. Another member (D219) of this clone, which was isolated from another patient at another location, was also sequenced and used for comparative analysis.

The penicillin binding proteins (PBP) PBP2x, PBP2b and PBP1a, which play an important role in penicillin resistance, were identical in all three genomes and unique compared to other *S. pneumoniae* except for PBP1a which was identified in one *S. pneumoniae* strain HMC3243. Most interestingly, a mosaic block of PBP2x was found in another isolate (*S. mitis* B93-4) from the same host, indicating horizontal gene transfer from *S. pneumoniae* to *S. mitis*. Amino acid changes associated with the PBP alleles of ST10523 agree with the intermediate penicillin resistance of this clone.

All pneumococcal major virulence factors were present in the ST1023 isolates and identical or highly similar to the laboratory strain *S. pneumoniae* R6, except for a *pspA* variant encoding a choline-binding protein. One remarkable difference was found in the hyaluronidase gene *hlyA*, which contains deletions within the promoter region and the coding region and thus appears to be non-functional.

In addition to almost 200 SNVs present in the *S. pneumoniae* D219 genome compared to those of *S. pneumoniae* D122 and D141, *S. pneumoniae* D219 contains a phage relict and a prophage

carrying at least two genes putatively involved in virulence. Furthermore, *S. pneumoniae* D219 carries a cluster of five genes, which is missing in the other two isolates. No ST10523 specific genes were found.

The presence of a non-functional hyaluronidase and the lack of the prophage containing putative virulence factors might contribute to the long persistence of *S. pneumoniae* D122/D141 strains in the CF patient.

Own contribution to the paper:

Assembly of sequence reads and final generation of genome sequences from assembled contigs including annotation of the three *S. pneumoniae* ST10523 isolates (D122, D141 and D219) and submission of genome sequences to the NCBI database. Comparative analysis of the three genomes including single nucleotide variation (SNV) retrieval and analysis of diverging sequence regions. Manual SNV retrieval and confirmation of results of automatically performed analysis by a newly developed wrapper tool. Detailed comparison of the serotype 23F capsule of the ST10523 genomes with the capsule cluster of *S. pneumoniae* ATCC 700669 (Acc. No. NC_011900; afterwards referred to as 23F) (Croucher, et al., 2009). Individual analysis of single virulence factors. Extraction of coding regions and deduction of proteins for determination of proteins specific to the ST10523 clone by comparison with other *S. pneumoniae* genomes. Unpublished work is described in chapter 3.1.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE
Clinical Science and Epidemiology

Long Persistence of a *Streptococcus pneumoniae* 23F Clone in a Cystic Fibrosis Patient

Martin Rieger,[a]* Harald Mauch,[b] Regine Hakenbeck[a]

Department of Microbiology, University of Kaiserslautern, Kaiserslautern, Germany[a]; HELIOS Klinikum Emil Von Behring, Berlin, Germany[b]

ABSTRACT *Streptococcus pneumoniae* isolates of serotype 23F with intermediate penicillin resistance were recovered on seven occasions over a period of 37 months from a cystic fibrosis patient in Berlin. All isolates expressed the same multilocus sequence type (ST), ST10523. The genome sequences of the first and last isolates, D122 and D141, revealed the absence of two phage-related gene clusters compared to the genome of another ST10523 strain, D219, isolated earlier at a different place in Germany. Genomes of all three strains carried the same novel mosaic penicillin-binding protein (PBP) genes, *pbp2x*, *pbp2b*, and *pbp1a*; these genes were distinct from those of other penicillin-resistant *S. pneumoniae* strains except for *pbp1a* of a Romanian *S. pneumoniae* isolate. All PBPs contained mutations that have been associated with the penicillin resistance phenotype. Most interestingly, a mosaic block identical to an internal *pbp2x* sequence of ST10523 was present in *pbp2x* of *Streptococcus mitis* strain B93-4, which was isolated from the same patient. This suggests interspecies gene transfer from *S. pneumoniae* to *S. mitis* within the host. Nearly all genes expressing surface proteins, which represent major virulence factors of *S. pneumoniae* and are typical for this species, were present in the genome of ST10523. One exception was the hyaluronidase gene *hlyA*, which contained a 12-nucleotide deletion within the promoter region and an internal stop codon. The lack of a functional hyaluronidase might contribute to the ability to persist in the host for an unusually long period of time.

IMPORTANCE *Streptococcus pneumoniae* is a common resident in the human nasopharynx. However, carriage can result in severe diseases due to a unique repertoire of pathogenicity factors that are rare in closely related commensal streptococci. We investigated a penicillin-resistant *S. pneumoniae* clone of serotype 23F isolated from a cystic fibrosis patient on multiple occasions over an unusually long period of over 3 years that was present without causing disease. Genome comparisons revealed an apparent nonfunctional pneumococcus-specific gene encoding a hyaluronidase, supporting the view that this enzyme adds to the virulence potential of the bacterium. The 23F clone harbored unique mosaic genes encoding penicillin resistance determinants, the product of horizontal gene transfer involving the commensal *S. mitis* as donor species. Sequences identical to one such mosaic gene were identified in an *S. mitis* strain from the same patient, suggesting that in this case *S. pneumoniae* played the role of donor.

KEYWORDS 23F clone, *Streptococcus pneumoniae*, cystic fibrosis, hyaluronidase, penicillin-binding proteins, persistence

Streptococcus pneumoniae is a common member of the commensal flora of the nasopharynx, particularly in children. Carriage rates between 5% and 20% have been observed in healthy children in Europe and the United States (1–3); however, high

Received 29 April 2017 **Accepted** 1 May 2017 **Published** 7 June 2017

Citation Rieger M, Mauch H, Hakenbeck R. 2017. Long persistence of a *Streptococcus pneumoniae* 23F clone in a cystic fibrosis patient. mSphere 2:e00201-17. https://doi.org/10.1128/mSphere.00201-17.

Editor Mariana Castanheira, JMI Laboratories

Address correspondence to Regine Hakenbeck, hakenb@rhrk.uni-kl.de.

* Present address: Martin Rieger, neox Aktiengesellschaft für Informationstechnologie, Pirmasens, Germany.

A Streptococcus pneumoniae 23F clone lacking a functional hyaluronidase was recovered from a cystic fibrosis patient over 37 months.

rates of over 80% are reported occasionally (4–6) and especially in developing countries (7, 8). Carriage may lead to a variety of diseases, such as otitis media, pneumonia, septicemia, and meningitis, especially in young children, elderly people, and immunocompromised patients (for a review, see reference 9). In fact, pneumococcal infections cause more deaths than other infectious diseases worldwide. The pathogenic potential distinguishes *S. pneumoniae* from other members of the group of viridans streptococci (10).

Numerous virulence factors of *S. pneumoniae* have been described (for a review, see references 11 and 12), but most of them are also present in the closely related commensal *Streptococcus* species *S. mitis*, *S. pseudopneumoniae*, and *S. oralis* (13–15). Typical for *S. pneumoniae* is the polysaccharide capsule, which is crucial for the pathogenicity of this species. Over 90 capsular types have been reported, based on biochemical and genetic analyses (16), and the potential to cause disease depends on the serotype (ST) (17). Before introduction of pneumococcal conjugate vaccines (PCVs), only a few serotypes accounted for the majority of invasive diseases, including types 4, 6B, 9V, 14, 18C, 19F, and 23F. The prevalence of serotypes changed after introduction of the first seven-valent conjugated vaccine (PCV7) in 2000, which covered serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F, followed later by PSV10, which also included serotypes 1, 5, and 7F, and by PCV13, with the additional serotypes 3, 6A, and 19A. Vaccination was accompanied by the appearance of antibiotic-resistant clones expressing nonvaccine serotypes (18). Other important virulence factors present in most *S. pneumoniae* strains are the pneumolysin Ply, choline-binding proteins (CBPs), which include the autolysin LytA as well as the variable CBPs PspA, PspC, and PspA, and the hyaluronidase HlyA (11, 12).

Due to its ability for genetic transformation, the genomes of *S. pneumoniae* isolates are highly diverse and include a large accessory genome. The increasing number of available genome sequences has provided an insight into the astounding repertoire of genes available in the pan-genome of pneumococcus (19, 20). The current standard for the definition of clones is based on comparative sequence analysis of housekeeping genes, which are part of the core genome common to all strains of the species; the methods is termed multilocus sequence typing (MLST) (21). The *Streptococcus pneumoniae* MLST database (https://pubmlst.org/spneumoniae) listed 13,126 STs in February 2017. Different capsular serotypes may be found within one ST due to capsule switching (22, 23). Genomes of an identical ST may vary considerably in their accessory genome content (24).

We report here on a rare clone of serotype 23F *S. pneumoniae* representing isolates with intermediate penicillin resistance which have been collected over a period of over 3 years from a patient in Berlin, Germany, with cystic fibrosis (CF); presence of the clone was not associated with disease. The genome sequences of three isolates of the same clone, including one isolate obtained from a different hospital in Germany, were used for comparative analysis of penicillin resistance determinants, the penicillin-binding proteins PBP2x, PBP1a, and PBP2b, and the main pneumococcal virulence factors.

RESULTS

Twenty-nine *S. pneumoniae* isolates were obtained between 1992 and 1995 from the Wannsee-Lungenklinik-Heckeshorn in Berlin. Seven of these isolates were recovered from one CF patient over a period of 37 months and were not associated with disease (Table 1). All seven strains expressed serotype 23F and showed identical antibiotic resistance patterns that were distinct from patterns of the other 22 strains (data not shown). MLST revealed that these seven isolates were members of the same clone of a new ST, ST10523. Screening of our strain collection detected another member of ST10523, strain D219, which was isolated in Leipzig, Germany, in 1989 (25). In order to see whether special virulence factors are associated with this clone and whether the isolates from the CF patient differed from D219, the genomes of the first (D122) and last (D141) isolate from Berlin and of D219 from Jena were sequenced.

Antibiotic resistance and penicillin-binding proteins. All seven ST10523 strains had intermediate resistance to beta-lactam antibiotics but were sensitive to tetracy-

TABLE 1 Bacterial strains[a]

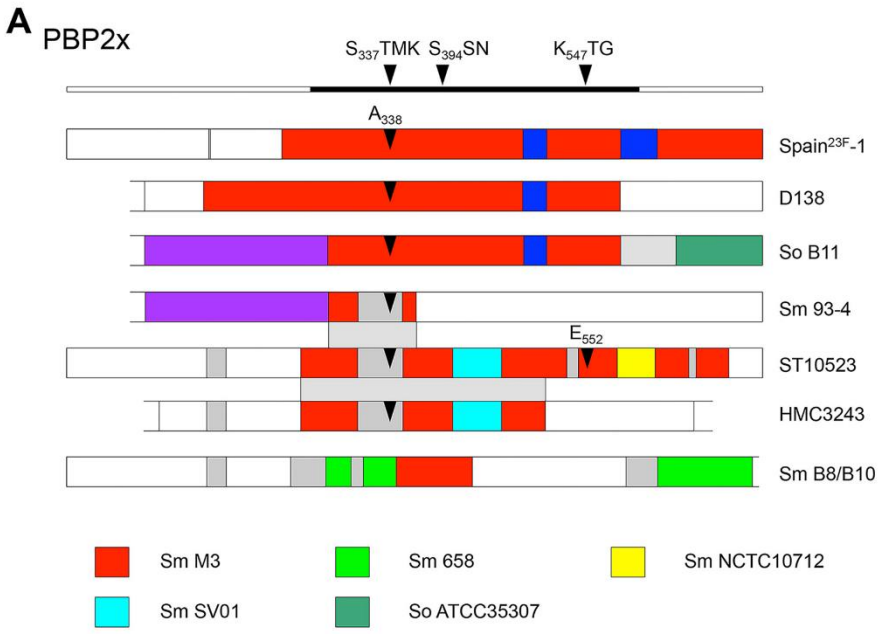
Species and isolate no.[b]	Date of isolation (day/mo/yr)	Site	ST	MIC (μg/ml)			TET/CLO/ERY susceptibility
				PEN-G	CTX	OXA	
S. pneumoniae (ST23F)							
D122	27/07/1992	Nasopharynx	10523	0.19–0.25	0.125–0.19	4–6	S/S/S
D127	8/4/1994	Nasopharynx	10523	0.19–0.25	0.125–0.19	4–6	S/S/S
D128	25/07/1994	Nasopharynx	10523	0.19–0.25	0.125–0.19	4–6	S/S/S
D134	31/10/1994	Nasopharynx	10523	0.19–0.25	0.125–0.19	4–6	S/S/S
D136	9/1/1995	Nasopharynx	10523	0.19–0.25	0.125–0.19	4–6	S/S/S
D139	9/5/1995	Sputum	10523	0.19–0.25	0.125–0.19	ND[d]	S/S/S
D141	1/8/1995	Nasopharynx	10523	0.19–0.25	0.125–0.19	4–6	S/S/S
D219	1989	Throat	10523	0.19–0.25	0.125–0.19	ND	S/S/S
S. mitis							
B8	1995	Oral cavity		0.12–0.2	0.006–0.03	0.5	S/S/S
B93-4[c]	1995	Oral cavity		ND	ND	ND	ND
B10	1995	Oral cavity		0.47	0.23	0.38	S/S/S
S. oralis							
B11	1995	Oral cavity		8	2	96	R/S/(R)

[a]S, sensitive; R, resistant; (R) intermediate resistant. Drug abbreviations: PEN-G, benzylpenicillin; CTX, cefotaxime; OXA, oxacillin; TET, tetracycline; CLO, chloramphenicol; ERY, erythromycin.
[b]All isolates were obtained from the Wannsee-Lungenklinik-Heckeshorn, Berlin, and from the same patient, except for D219, which was isolated in Leipzig (25).
[c]The strain could not be recovered after DNA isolation.
[d]ND, not determined.

cline, erythromycin, and chloramphenicol (Table 1). Since PBP2x, PBP2b, and PBP1a play key roles in penicillin resistance, these genes were analyzed in detail to see whether they are related to PBP alleles of other penicillin-resistant *S. pneumoniae* (PRSP) strains. The PBP sequences of D219, D122, and D141 were identical and contained sequence blocks that diverged from PBP genes of the penicillin-sensitive R6 laboratory strain by ~20%. They did not match any other PBP sequences in the NCBI database except for *pbp1a* (see below). Interestingly, the genomic regions containing *pbp2x* (spr302 to spr307; *yllC* to *clpL*) and *pbp2b* (spr1513 to spr1517; *mutT* to *pbp2b*) contained a significantly larger amount of single nucleotide polymorphisms (SNPs; >5.3%) in the ST10523 strains than in the entire genome of the R6 strain, excluding variable genes (<0.7%), indicating transfer of large sequence blocks flanking the PBP genes. Similar observations have been reported previously (26). In contrast, flanking genes of *pbp1a* showed no signs of recombination events.

The PBP2x gene of ST10523 has a complex mosaic structure (Fig. 1). It contains sequence blocks that are highly related (<3% difference on the DNA level) to *pbp2x* of putative penicillin-sensitive donor *S. mitis* strains M3, SV01, and NCTC10712 (27), in addition to other divergent sequences of unknown origin (Fig. 1A). The deduced protein sequence included mutations A_{338} and E_{552}, which are known to confer resistance to beta-lactams; both amino acid changes are frequent in clinical PRSP isolates. The only other 2 amino acids that did not match any of the sensitive reference *S. mitis* strains were T_{410}, present in *pbp2x* of low-level-resistant viridans streptococci, and T_{434}, which occurs in some penicillin-sensitive strains (Fig. 1B) (27).

Furthermore, we obtained *pbp2x* sequences from four commensal streptococcus isolates from the same patient, isolates D122 and D141 (Table 1). The PBP2x genes of *S. mitis* strains B8 and B10 were identical and were distinct from *pbp2x* of *S. oralis* B11 and that of *S. pneumoniae* ST10523. The gene $pbp2x_{B11}$ contained a central sequence block almost identical to that of Spain^{23F}-1, and 3′-sequences were related to *S. oralis* ATCC 35307 (Fig. 1A). BLAST searches revealed one Romanian isolate, HMC3243, of unknown serotype (28) which was partially identical to D219 up to codon 517, whereas the C-terminal part represented R6 sequences (Fig. 1A and B). Interestingly, *pbp2x* of *S. mitis* B93-4 contained a 284-nucleotide (nt) sequence block almost identical to $PBP2x_{D219}$ (codons 284 to 377), including two SNPs resulting in 1 amino acid change, $D_{368}N$, suggesting that this block was acquired from *S. pneumoniae* ST10523 (Fig. 1A).



B

```
                222222333333333333333333333333444444444444444455555555555555555555555555555566666
                677881112344455566666667788889901133446688889990001111123333345666666677779900111
                858121680836715801346891824594810704479213680181560346783156786223456782346784524 6
R6              PTDQEDEIETMAANGVNSELIDAIEGRMSHVTANTANQSISIPDTAINKDLTNVLVVSPTVTLQLVGLTDYAVSSGIANSMD
D219            ...L.ND.KAT.S..Y........DTG....ST..S.MAL..TNSV.KE....I..LY.II..E...S.N.V.TN......E
HMC32432        ...L.ND.KAT.S..Y........DTG....ST..S.MAL..TNSV.KE....I.............................
Sm b93-4        .S...ND.KAT.S..Y.....N..........................................................
138             ...L.N...ATSS.SY...F...T.TG.L..S.K..S..L..TNSV...ETNHI..LY.II.V....S.N...TN...T...
23F             ...L.N...ATSS.SY...F...TDTG.L..S.K..S..L..TNSV...ETNHI..LY.II.V....S.N.V.TN...T..E
Sm B8/B10       T.N....LK.T....Y......V.DT..LL.S.....................I.I.............N....N....V..
So B11          .S...N...ATSS.SY...F...TDTG.L..S.K..S..L..TNSV...ETNHI..LY.II.V....S.N.V.TN...T..E
Sm M3           ...L.N....TSS.SY........DT.....S.......L..TNSV.KE....I..LY.II.V....S.N.V.TNAAS...E
Sm NCTC10712    ...LQN....TSS..Y..........G...I.........TL...V....Q.H...R.A..N.....S.N.V.AN......E
Sm SV01         .....ND.K....K.YTR..T....SGT..I....T.MAL.L..SVV..EVDH.M.T.A..N...T.S.N.V..H......E
So ATCC35037    .....ND.K..S...Y.........SGT..I...AT.M.L.L..SV...ASEH.M....N.N...T.E.N.V..H.V...TE
```

FIG 1 Schematic representation of the mosaic genes for PBP2x, PBP1a, and PBP2b from *S. pneumoniae* and *S. mitis* strains. (A) Mosaic sequence blocks in PBP2x. PBP2x genes of ST10523 (*S. pneumoniae* strains D122, D139, D141, and D219) are identical; *pbp2x* of *S. pneumoniae* Spain^{23F}-1 is included for comparison. The *pbp2x* gene of *S. pneumoniae* HMC3243 from Romania, which is partially identical to *pbp2x* of ST10523, was also included. The colors indicate closely related sequence portions with >95% identity to *pbp2x* reference sequences from penicillin-sensitive *S. mitis* (Sm.) strains M3, 658, NCTC10712, and SV01 and *S. oralis* (So) ATCC 35307 (76), as indicated by the color coded boxes. White indicates sequences related to the penicillin-sensitive strain *S. pneumoniae* R6, and gray sequences are highly divergent from R6 sequences and other *pbp2x* reference sequences. The domain structure of PBP2x with the three homology boxes is shown above; the central penicillin-binding domain is indicated by the black bar. Black arrows represent mutations A_{338} and E_{552} in *pbp2x*. The regions that are identical between *pbp2x* of ST10523, *S. mitis* B93-4, and *S. pneumoniae* HMC3243 are indicated by the shaded area. (B) Deduced amino acid sequences of the transpeptidase domain of PBP2x; PBP2x of *S. pneumoniae* R6 was used as a reference. The positions of the amino acids are indicated vertically in the top three rows. Only positions with altered residues are shown; residues identical to those in strain R6 are indicated by dots. The mutations A_{338} and E_{552} are highlighted by black arrowheads; white arrowheads indicate amino acids which are not present in PBP2x of the penicillin-sensitive reference strains (see text for details).

The flanking 5′-sequence block was identical to *S. oralis* B11, and the 3′-end was identical to the *S. pneumoniae* R6 sequence, documenting multiple interspecies gene transfer events.

The internal mosaic blocks of ST10523 *pbp2b* and *pbp1a* were highly related to the respective genes in *S. mitis* NCTC10712 (see Fig. S1 in the supplemental material).

PBP2b contained the mutation A_{446} close to the conserved S_{443}SN motif, which mediates low resistance levels (29) and which is present in most penicillin-resistant isolates. Moreover, it had one more amino acid, Y_{430}, that resulted in a deduced protein of 681 residues (Fig. S1). Regarding PBP1a of ST10523, the change of four consecutive residues, T_{574}SQF to NTGY, has been associated with penicillin resistance, but the mutation A_{371} within the active site motif S_{470}TMK, which also has been implicated in penicillin resistance, was not present (30–32). The PBP1a gene of ST10523 was identical to that of *S. pneumoniae* strain HMC3243 (Fig. S1) except for three silent SNPs. In contrast, the mosaic structure of *pbp2b* was entirely different. This clearly indicated that the three PBP genes were acquired from different sources or occasions in *S. pneumoniae* HMC3243 compared to ST10523. In summary, all three PBPs contained amino acid mutations that have been associated with the resistance phenotype corresponding to the intermediate penicillin resistance of ST10523.

Genomic comparison of *S. pneumoniae* genomes. Since ST10523 is a new sequence type, the 2,050,063-nt draft genome of strain D219 was first compared to the R6 genome. Genes with no match in R6 were then used in a BLAST search of the NCBI database. Excluding transposases, the genome of D219 differed from the R6 genome by ~8%, including parts of a bacteriocin cluster (SPND219_00557 to SPND219_00567), the *cps* biosynthesis cluster encoding the 23F capsule (SPND219_00380 to SPND219_00398), and two clusters encoding phage-related genes (SPND219_00003 to SPND219_00023 [phage relict] and SPND219_01526 to SPND219_01585 [prophage]). This percentage corresponds to data obtained in comparative genomic hybridization on an oligonucleotide microarray representing the TIGR4 genome (33). Large parts of the phage relict were present in several genomes of *S. pneumoniae*. The prophage shows high similarity to *S. pneumoniae* phage 040922 (GenBank accession number FR671406), which is associated with a Tn*916*-like element in one *S. pneumoniae* strain, 18C/3 (34). The prophage contains two large genes (fragments SPND219_01501-3 and SPND219_01535) that encode the surface-expressed tail fiber PblB and the tape measure protein PblA. Homologues of PblA and PblB of *S. mitis* phage SM1 have been shown to be involved in the platelet-binding activity of *S. mitis* SF100 (35) and for its virulence in an animal model of infective endocarditis (36). No genes exclusively carried by ST10523 could be detected in BLAST searches.

Between the genome of D219 on one hand and D122 and D141 on the other hand, little difference regarding gene content was noted. The only exceptions were the phage relict and the prophage mentioned above, which were absent in D122 and D141. One gene cluster, SPND122_00705 to SPND122_00709 and SPND141_00707 to SPND141_00711 related to the R6 genes *spr0623* to *spr0627* (ABC transporter, lactate monooxygenase, 2-lysyl-tRNA synthetase) were not found in D219. However, since these genes are all located on small contigs, including repeat elements such as BOX and RUP (37, 38), which result in problems during the genome assembly process, verification of their absence in D219 will require further analyses.

The genomes of D122 and D141 differed from each other by approximately 0.01% SNPs in 49 genes, less than that found in D219 (0.02 to 0.024% affecting a total of 178 genes in D219), i.e., the isolates from the patient are more closely related to each other than to D219, in agreement with their distinct place of isolation. One clear difference was observed in the gene encoding the two-component sensor kinase HK07 (39), which is not essential in *S. pneumoniae* (40). It was intact in strains D122 and D219 (SPND122_00180 and SPND219_00200), whereas an ISL3 family transposase fragment was inserted into the D141 gene, resulting in three incomplete gene fragments, SPND141_00182 to SPND141_00184.

Virulence genes in ST10523. All genes encoding the main *S. pneumoniae* specific virulence factors *lytA*, *ply*, *Nana*, and *nanB* and the CBPs *pspC*, *pspA*, and *pcpC* were present in the three ST10523 genomes. The deduced protein sequences were identical to each other and identical or highly similar to R6 proteins except for *pspA*, which represented a distinct genetic variant in ST10523. Differences in the repeat regions in

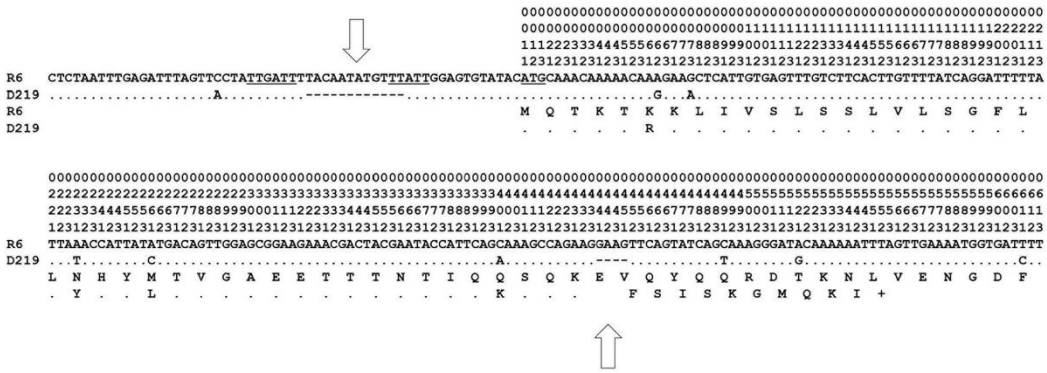


FIG 2 The HlyA gene of *S. pneumoniae* R6 and ST10523. The HlyA gene of *S. pneumoniae* R6, including upstream sequences (57 nt) to codon 72, is indicated. The −35 region, the −10 region, and the start codon ATG are underlined. The deduced amino acid sequence is indicated below. In the *hlyA* region of D219 (ST10523), only nucleotides and deduced amino acids that differ from the reference R6 sequence are shown. The 12-nt deletion upstream of the ATG start codon and the 4-nt deletion in the D219 sequence are indicated by open arrows. Vertical numbers in the first three rows refer to codons; numbers 1, 2, and 3 in the fourth row indicate the first, second, and third positions in the respective codon.

genes encoding CBPs were not considered, since they are most likely caused by assembly problems of the choline-binding repeat regions. Other genes encoding surface proteins of important biological function, *pavA* and, for the IgA proteases, *zmpB* and *zmpA* were also present and identical in all three strains, except for the IgA protease, which contained gaps in the sequence of D219 and two SNPs plus a single-nucleotide deletion in D122, resulting in a premature stop codon. No pilus cluster was detectable. The capsule cluster SPND219_00380 to SPND219_00394 differed from that of *S. pneumoniae* strain Spain^{23F}-1 by 20 SNPs, as expected for genes of distinct clonal origins. The last four-gene *rml* operon encoding enzymes involved in dTDP-rhamnose synthesis and which is common to several serotypes (41) included two short divergent regions in *rmlB* identical to *rmlB* of many serotype 19F strains.

However, all three ST10523 genomes contained significant differences in *hlyA*, which encodes the hyaluronidase (SPND219_00348, SPND122_00329, and SPND141_00331) and which are unusual in other *S. pneumoniae* genomes. Within the HlyA gene, a 4-bp deletion corresponding to the position 128 nt downstream of the putative ATG start codon of R6 resulted in a premature stop codon. Moreover, a 12-nt deletion within the promoter region 10 nt upstream of the ATG start codon was present (Fig. 2). This strongly suggests that no functional hyaluronidase is expressed in ST10523 strains. Searches in the whole-genome contig NCBI database revealed another six genomes which contained this peculiarity (Table 2), but they differed by up to four SNPs from the D219 region shown in Fig. 2.

DISCUSSION

Prolonged carriage of the same *S. pneumoniae* clone for a period of 37 months, as observed for serotype 23F isolates obtained from a CF patient in our study, is unusual.

TABLE 2 *S. pneumoniae* genome sequences with an incomplete *hlyA*

Accession no.	Strain	Serotype	Source	Date (day/mo/yr)	Country (region or city)
LJVO01000185	NTPn 4	NT	Blood	2004	South Africa (KwaZulu-Natal)
CVHP01000011	0338	NT	Blood	2001	USA (Alaska)
CKDL01000005	Type strain	NT	Nasopharynx	14/4/2008	Thailand (Maela)
CPLS01000001	LMG205	6B	Not known	2008	Thailand
CRPU01000001	SMRU824	NT	Nasopharynx	10/10/2008	Thailand (Maela)
AGOE01000004	GA16531	NA[a]	NA	2001	USA (metropolitan Atlanta)
CFNW01000018	6378-99	19F	Not known	1999	USA (Tennessee)

[a]NA, not available.

S. pneumoniae is not considered a persistent colonizer in CF patients, unlike *Pseudomonas aeruginosa* and *Staphylococcus aureus* (42–44). Long-term persistence has been reported for *Staphylococcus aureus* for up to 70 months (45). The duration of pneumococcal carriage in healthy children is a few weeks, ranging from 2 days and in rare cases up to 6 to 12 months, depending on the age of the carrier and the serotype of the isolate (2, 3, 46–48). Some serogroups, including serotype 23F, are generally carried for longer periods than other serogroups (46, 49) and have a low propensity to cause invasive disease (50). Interestingly, an inverse relationship between the attack rate of a given capsular serotype and its duration of carriage has been noted (3).

Carriage rates for *S. pneumoniae* isolated from CF patients are similar, with colonization rates ranging between 3 and 20% (51–56). No special serotypes appear to be associated with CF (57), but some serotypes may be more common, depending on the geographic area. Serotype 23F isolates were prevalent in a CF unit in Madrid, mainly due to the clone Spain^{23F}-1 of a varied multiresistance phenotype (52). Serotype 3 prevailed in another study which reported no 23F serotype isolates in a CF center in Rome, probably because all patients had received vaccination (58). In both studies, *S. pneumoniae* was recovered more than once from some patients. However, only three strains of serotype 23F isolated over a period of 3 months showed identical SmaI restriction patterns, revealed by pulsed-field gel electrophoresis (PFGE), and the same antibiotic resistance profile, and thus most likely represented members of the same clone (52). Three patients carried *S. pneumoniae* with the same serotype and identical SmaI PFGE pattern for 1 to 8 months (58). In these cases, *S. pneumoniae* was considered a colonizer, since at the time of isolation the patients showed no evidence of pulmonary exacerbation.

Genomic comparisons showed that the two strains, D122 and D141, from the CF patient are more closely related to each other than to D219, which was isolated 3 years before from a different geographical site. They differed from D219 by the absence of two large gene clusters encoding a prophage and a phage relict, by the presence of a five-gene cluster, including an amino acid ABC transporter, and by the estimated number of SNPs. The prophage carries two genes encoding large proteins PblA and PblB. Homologues of these proteins are frequent in *S. pneumoniae* phages (59) and have been shown to play a role in adhesion and virulence in *S. mitis* (35, 36). It these proteins play a similar role in *S. pneumoniae*, it is conceivable that their absence in D122 and D141 supports extended carriage.

The variation between D122 and D141 concerned only 49 genes, and D141 contained an insertion of an ISL3 family transposase fragment into the gene encoding the histidine protein kinase HK07, which was absent in the D122 gene. This element was present at another three sites in the D141 genome and at one site in the D122 genome, whereas it could not be detected in D219. Other studies have supported little genomic variation during carriage of the same *S. pneumoniae* clone. Minimal variation was observed during carriage established experimentally with a single serotype 6B strain of ST138 (60). The maximum SNP distance between any of the 229 isolates obtained over a period of 35 days versus the reference strain was three SNPs (60). It should be noted that the genomic comparison between two isolates of strain D39, a historically important serotype 2 isolate from the early 1940s (61) and which have been cultivated separately for at least 21 years, revealed only five mutations (62). Similarly, some strains isolated during a 7-month period from a child with chronic pneumococcal infection varied by only ≤30 SNPs (63). However, those authors noted there were also multiple events of horizontal gene transfer in some strains, which most likely occurred during polyclonal infection. In contrast, we saw no evidence of gene acquisition in D141 versus D122, and the *S. mitis* strain B93-4 contained a *pbp2x* fragment identical to *pbp2x* of ST10523. The mosaic PBP2x and PBP2b genes of ST10523 represent new gene variants and are distinct from all others found in the NCBI database. Therefore, this finding indicates interspecies gene transfer from *S. pneumoniae* to *S. mitis* in the same host.

No genes specifically associated with ST10523 genomes were identified. This is not astounding, given the vast number of actually available genomes of *S. pneumoniae*.

Based on a pan-genome analysis of 158 *S. pneumoniae* genomes, it has been predicted that only 0.3 new genes will be discovered in a new genome if a data set from 1,000 genomes is already available, and only 0.06 new genes will be discovered from 5,000 genomes (24). However, an unusual hyaluronidase gene, *hlyA*, was present in the ST10523 genomes which contained a stop codon (Fig. 2) distinct from that described in a serotype 3 clone, ST180, where an SNP at position 376 of the hyaluronidase coding sequence resulted in a stop codon and truncation of the protein after 125 amino acids (24). Moreover, a 12-nt deletion within the promoter region 10 nt upstream of the ATG was detected. Hyaluronidase is produced by almost all clinical isolates of pneumococci (64). It is one of the genes which have not been found in closely related viridans streptococci, except for some *S. oralis* isolates, but not in *S. mitis* strains (65), i.e., it represents a typical component of the species *S. pneumoniae*. The enzyme depolymerizes hyaluronic acid, which is an important component of the host connective tissue and extracellular matrix (66). No significant impact on virulence was found in a mouse intraperitoneal infection model when a single *hlyA* mutant was used (67); however, when a double mutant that was also deficient for the pneumolysin gene *ply* was tested, virulence was significantly decreased (68). Hyaluronidase significantly potentiated pneumolysin-mediated ciliary slowing and epithelial damage in an *in vitro* model, suggesting that its presence favors colonization and subsequently extrapulmonary dissemination of the pneumococcus (69). It is therefore tempting to assume that the lack of a functional hyaluronidase predestines ST10523 to survive within the host for an extended period without causing disease. In conclusion, the unusually long carriage rate observed for *S. pneumoniae* isolates D122 and D141 might not be related to mutations or genetic variants acquired during persistence in the human host, but rather to the loss of a large prophage carrying potential virulence factors and to the absence of a complete hyaluronidase gene.

MATERIALS AND METHODS

Bacterial strains and growth conditions. *S. pneumoniae* strains D122 to D141, *S. mitis* B8, B93-4, and B10, and *S. oralis* B11 (Table 1) were part of a strain collection obtained from the Wannsee-Lungenklinik-Heckeshorn in Berlin; strain D219, isolated in Leipzig, has been described elsewhere (25). Strains were grown at 37°C without aeration in complex C medium (70) supplemented with 0.1% yeast extract (C + Y). MICs were determined by the agar dilution method (for beta-lactams) or by using E-test strips for all other antibiotics (Oxoid GmbH, Basingstoke, United Kingdom) on D-agar plates supplemented with 3% sheep blood (71).

DNA sequencing and analysis. Chromosomal DNA from streptococci was isolated as described previously (72). Internal sequences of the seven housekeeping genes were obtained with primers described on the *Streptococcus pneumoniae* MLST homepage (https://pubmlst.org/spneumoniae). PBP2x gene fragments were amplified with the primers pn2xup and pn2xdown (72), and direct sequencing of PCR products was performed with consecutive primers. PCR products were purified using a JetQuick DNA purification kit (GenoMed). PCRs were performed using either Goldstar Red *Taq* polymerase (Eurogentec) or DreamTaq polymerase (Fermentas), according to the manufacturer instructions. The genomes of D219, D122, and D144 were sequenced using a 454 Life Sciences FLX sequencer, and reads were assembled by the 454 Newbler Assembler version 2.6. Contigs were aligned to the *S. pneumoniae* R6 genome sequence (73). The rapid annotation subsystem technology (RAST) server (74) designed for annotation of bacterial and archaeal genomes was applied to obtain EMBL-formatted files containing protein, tRNA, and rRNA annotations from a large set of several output formats.

DNA analysis and bioinformatic tools. For the analysis of SNPs in the three ST10523 genomes, only sequences that were 350 nt from contig ends were included, to avoid potential errors generated by the 454-generated sequences. Individual open reading frames were investigated manually and compared with other genome sequences by using BLAST analyses and the NCBI database (nucleotides and whole-genome contigs). As a reference for the serotype 23F capsule cluster, genes of strain ATCC (Spain^{23F}-1) were used (75). Alignments were prepared using Clustal X2 (76). Codon sites were included manually and trimmed by using the program Clustal Formatter 3 (http://nbc11.biologie.uni-kl.de/sequence_analysis/ClustalFormatter3/documentation.html) to reveal only sites that differ from the reference sequence shown in Fig. 1B.

Accession number(s). The following sequences have been submitted to GenBank and assigned the following accession numbers (shown in parentheses): for genomes, D219 (CP016227), D122 (CP016632), D141 (CP016633); for PBP genes of *S. pneumoniae* strain HMC3243, *pbp2x* (FJ439546), *pbp1a* (FJ439538), and *pbp2b* (FJ439554) (28); for PBP2x genes, *S. mitis* strain B8 (KY292528), strain SV01 (KY292540), and strain B93-4 (KY783589), and *S. oralis* strain B11 (KY783587).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSphere.00201-17.

FIG S1, TIF file, 6.6 MB.

ACKNOWLEDGMENTS

We thank Shwan Rachid for help with the MLST analysis and Jennifer Loewe for DNA sequencing of PBP genes.

This work was supported by the Bundesministerium für Bildung und Forschung (grant 0313801 1) and the Deutsche Forschungsgemeinschaft (grant Ha 1011/11-3 to R.H.).

REFERENCES

1. Short KR, Diavatopoulos DA. 2015. Nasopharyngeal colonization with *Streptococcus pneumoniae*, p 279–291. *In* Brown J, Hammerschmidt S, Orihuela C (ed), *Streptococcus pneumoniae* molecular mechanisms of host-pathogen interactions. Academic Press, London, United Kingdom.
2. Högberg L, Geli P, Ringberg H, Melander E, Lipsitch M, Ekdahl K. 2007. Age- and serogroup-related differences in observed durations of nasopharyngeal carriage of penicillin-resistant pneumococci. J Clin Microbiol 45:948–952. https://doi.org/10.1128/JCM.01913-06.
3. Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, Moxon ER, Crook DW, Peto TE. 2006. Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. J Infect Dis 194:682–688. https://doi.org/10.1086/505710.
4. Nunes S, Sá-Leão R, Carriço J, Alves CR, Mato R, Avô AB, Saldanha J, Almeida JS, Sanches IS, de Lencastre H. 2005. Trends in drug resistance, serotypes, and molecular types of *Streptococcus pneumoniae* colonizing preschool-age children attending day care centers in Lisbon, Portugal: a summary of 4 years of annual surveillance. J Clin Microbiol 43:1285–1293. https://doi.org/10.1128/JCM.43.3.1285-1293.2005.
5. Ercibengoa M, Arostegi N, Marimón JM, Alonso M, Pérez-Trallero E. 2012. Dynamics of pneumococcal nasopharyngeal carriage in healthy children attending a day care center in northern Spain. Influence of detection techniques on the results. BMC Infect Dis 12:69. https://doi.org/10.1186/1471-2334-12-69.
6. Wyllie AL, Chu ML, Schellens MH, van Engelsdorp Gastelaars J, Jansen MD, van der Ende A, Bogaert D, Sanders EA, Trzciński K. 2014. *Streptococcus pneumoniae* in saliva of Dutch primary school children. PLoS One 9:e102045. https://doi.org/10.1371/journal.pone.0102045.
7. Jebaraj R, Cherian T, Raghupathy P, Brahmadathan KN, Lalitha MK, Thomas K, Steinhoff MC. 1999. Nasopharyngeal colonization of infants in southern India with *Streptococcus pneumoniae*. Epidemiol Infect 123:383–388. https://doi.org/10.1017/S0950268899003131.
8. Hill PC, Cheung YB, Akisanya A, Sankareh K, Lahai G, Greenwood BM, Adegbola RA. 2008. Nasopharyngeal carriage of *Streptococcus pneumoniae* in Gambian infants: a longitudinal study. Clin Infect Dis 46:807–814. https://doi.org/10.1086/528688.
9. Henriques-Normark B, Tuomanen EI. 2013. The pneumococcus: epidemiology, microbiology, and pathogenesis. Cold Spring Harb Perspect Med 3:a010215. https://doi.org/10.1101/cshperspect.a010215.
10. Bogaert D, de Groot R, Hermans PW. 2004. *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. Lancet Infect Dis 4:144–154. https://doi.org/10.1016/S1473-3099(04)00938-7.
11. Mitchell TJ. 2003. The pathogenesis of streptococcal infections: from tooth decay to meningitis. Nat Rev Microbiol 1:219–230. https://doi.org/10.1038/nrmicro771.
12. Hammerschmidt S. 2007. Pneumococcal virulence factors and adhesion proteins targeting the host, p 141–203. *In* Hakenbeck R, Chhatwal GS (ed), Molecular biology of streptococci. Horizon Press, Wymondham, Norfolk.
13. Denapaite D, Brückner R, Nuhn M, Reichmann P, Henrich B, Maurer P, Schähle Y, Selbmann P, Zimmermann W, Wambutt R, Hakenbeck R. 2010. The genome of *Streptococcus mitis* B6: what is a commensal? PLoS One 5:e9426. https://doi.org/10.1371/journal.pone.0009426.
14. Shahinas D, Thornton CS, Tamber GS, Arya G, Wong A, Jamieson FB, Ma JH, Alexander DC, Low DE, Pillai DR. 2013. Comparative genomic analyses of *Streptococcus pseudopneumoniae* provide insight into virulence and commensalism dynamics. PLoS One 8:e65670. https://doi.org/10.1371/journal.pone.0065670.
15. Denapaite D, Rieger M, Köndgen S, Brückner R, Ochigava I, Kappeler P, Mätz-Rensing K, Leendertz F, Hakenbeck R. 2016. Highly variable *Streptococcus oralis* strains are common among viridans streptococci isolated from primates. mSphere 1:e00041-15. https://doi.org/10.1128/mSphere.00041-15.
16. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA, Samuel G, Skovsted IC, Kaltoft MS, Barrell B, Reeves PR, Parkhill J, Spratt BG. 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2:e31. https://doi.org/10.1371/journal.pgen.0020031.
17. Hausdorff WP, Feikin DR, Klugman KP. 2005. Epidemiological differences among pneumococcal serotypes. Lancet Infect Dis 5:83–93. https://doi.org/10.1016/S1473-3099(05)01280-6.
18. Torres A, Bonanni P, Hryniewicz W, Moutschen M, Reinert RR, Welte T. 2015. Pneumococcal vaccination: what have we learnt so far and what can we expect in the future? Eur J Clin Microbiol Infect Dis 34:19–31. https://doi.org/10.1007/s10096-014-2208-6.
19. Muzzi A, Donati C. 2011. Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae*. Int J Med Microbiol 301:619–622. https://doi.org/10.1016/j.ijmm.2011.09.008.
20. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER, Masignani V. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11:R107. https://doi.org/10.1186/gb-2010-11-10-r107.
21. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95:3140–3145. https://doi.org/10.1073/pnas.95.6.3140.
22. Brueggemann AB, Pai R, Crook DW, Beall B. 2007. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. PLoS Pathog 3:e168. https://doi.org/10.1371/journal.ppat.0030168.
23. Crisafulli G, Guidotti S, Muzzi A, Torricelli G, Moschioni M, Masignani V, Censini S, Donati C. 2013. An extended multi-locus molecular typing schema for *Streptococcus pneumoniae* demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. Infect Genet Evol 13:151–161. https://doi.org/10.1016/j.meegid.2012.09.008.
24. Tettelin H, Chancey S, Mitchell T, Denapaite D, Schähle Y, Rieger M, Hakenbeck R. 2015. Genomics, genetic variation, and regions of differences, p 81–107. *In* Brown J, Hammerschmidt S, Orihuela C (ed), *Streptococcus pneumoniae* molecular mechanisms of host-pathogen interactions. Academic Press, London, United Kingdom.
25. Reichmann P, Varon E, Günther E, Reinert RR, Lüttiken R, Marton A, Geslin P, Wagner J, Hakenbeck R. 1995. Penicillin-resistant *Streptococcus pneumoniae* in Germany: genetic relationship to clones from other European countries. J Med Microbiol 43:377–385. https://doi.org/10.1099/00222615-43-5-377.
26. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, Turner P, Bentley SD, Parkhill J. 2014. Comprehensive identification of single nucleotide poly-

morphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet 10:e1004547. https://doi.org/10.1371/journal.pgen.1004547.
27. van der Linden M, Otten J, Bergmann C, Latorre C, Linares J, Hakenbeck R. 2017. PBP2x in *Streptococcus pseudopneumoniae*: an insight into the diversity of PBP2x alleles and mutations in viridans streptococci. Antimicrob Agents Chemother 61:e02646-16. https://doi.org/10.1128/AAC.02646-16.
28. Kosowska-Shick K, McGhee P, Appelbaum PC. 2009. Binding of faropenem and other beta-lactam agents to penicillin-binding proteins of pneumococci with various beta-lactam susceptibilities. Antimicrob Agents Chemother 53:2176–2180. https://doi.org/10.1128/AAC.01566-08.
29. Grebe T, Hakenbeck R. 1996. Penicillin-binding proteins 2b and 2x of *Streptococcus pneumoniae* are primary resistance determinants for different classes of β-lactam antibiotics. Antimicrob Agents Chemother 40:829–834.
30. Smith AM, Klugman KP. 2003. Site-specific mutagenesis analysis of PBP 1A from a penicillin-cephalosporin-resistant pneumococcal isolate. Antimicrob Agents Chemother 47:387–389. https://doi.org/10.1128/AAC.47.1.387-389.2003.
31. Smith AM, Klugman KP. 1998. Alterations in PBP 1A essential for high-level penicillin resistance in *Streptococcus pneumoniae*. Antimicrob Agents Chemother 42:1329–1333.
32. Job V, Di Guilmi AM, Martin L, Vernet T, Dideberg O, Dessen A. 2003. Structural studies of the transpeptidase domain of PBP1a from *Streptococcus pneumoniae*. Acta Crystallogr D Biol Crystallogr 59:1067–1069. https://doi.org/10.1107/S0907444903006954.
33. Hakenbeck R, Balmelle N, Weber B, Gardès C, Keck W, de Saizieu A. 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. Infect Immun 69:2477–2486. https://doi.org/10.1128/IAI.69.4.2477-2486.2001.
34. Wyres KL, van Tonder A, Lambertsen LM, Hakenbeck R, Parkhill J, Bentley SD, Brueggemann AB. 2013. Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated prior to 1974. BMC Genomics 14:500. https://doi.org/10.1186/1471-2164-14-500.
35. Bensing BA, Rubens CE, Sullam PM. 2001. Genetic loci of *Streptococcus mitis* that mediate binding to human platelets. Infect Immun 69:1373–1380. https://doi.org/10.1128/IAI.69.3.1373-1380.2001.
36. Mitchell J, Siboo IR, Takamatsu D, Chambers HF, Sullam PM. 2007. Mechanism of cell surface expression of the *Streptococcus mitis* platelet binding proteins PblA and PblB. Mol Microbiol 64:844–857. https://doi.org/10.1111/j.1365-2958.2007.05703.x.
37. Oggioni MR, Claverys JP. 1999. Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. Microbiology 145:2647–2653. https://doi.org/10.1099/00221287-145-10-2647.
38. Martin B, Humbert O, Càmara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R, Morrison DA, Boulnois GJ, Claverys J-P. 1992. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res 20:3479–3483. https://doi.org/10.1093/nar/20.13.3479.
39. Lange R, Wagner C, de Saizieu A, Flint N, Molnos J, Stieger M, Caspers P, Kamber M, Keck W, Amrein KE. 1999. Domain organization and molecular characterization of 13 two-component systems identified by genome sequencing of *Streptococcus pneumoniae*. Gene 237:223–234. https://doi.org/10.1016/S0378-1119(99)00266-8.
40. Throup JP, Koretke KK, Bryant AP, Ingraham KA, Chalker AF, Ge Y, Marra A, Wallis NG, Brown JR, Holmes DJ, Rosenberg M, Burnham MK. 2000. A genomic analysis of two-component signal transduction in *Streptococcus pneumoniae*. Mol Microbiol 35:566–576. https://doi.org/10.1046/j.1365-2958.2000.01725.x.
41. Morona JK, Miller DC, Coffey TJ, Vindurampulle CJ, Spratt BG, Morona R, Paton JC. 1999. Molecular and genetic characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 23F. Microbiology 145:781–789. https://doi.org/10.1099/13500872-145-4-781.
42. Renders N, Verbrugh H, Van Belkum A. 2001. Dynamics of bacterial colonisation in the respiratory tract of patients with cystic fibrosis. Infect Genet Evol 1:29–39. https://doi.org/10.1016/S1567-1348(01)00004-1.
43. Iacocca VF, Sibinga M, Barbero GJ. 1963. Respiratory tract bacteriology in cystic fibrosis. Am J Dis Child 106:315–324. https://doi.org/10.1001/archpedi.1963.02080050317012.
44. Høiby N. 1982. Microbiology of lung infections in cystic fibrosis patients. Acta Paediatr 71:33–54. https://doi.org/10.1111/j.1651-2227.1982.tb09640.x.
45. Kahl BC, Duebbers A, Lubritz G, Haeberle J, Koch HG, Ritzerfeld B, Reilly M, Harms E, Proctor RA, Herrmann M, Peters G. 2003. Population dynamics of persistent *Staphylococcus aureus* isolated from the airways of cystic fibrosis patients during a 6-year prospective study. J Clin Microbiol 41:4424–4427. https://doi.org/10.1128/JCM.41.9.4424-4427.2003.
46. Gray BM, Converse GM, III, Dillon HC, Jr. 1980. Epidemiologic studies of *Streptococcus pneumoniae* in infants: acquisition, carriage, and infection during the first 24 months of life. J Infect Dis 142:923–933. https://doi.org/10.1093/infdis/142.6.923.
47. Ekdahl K, Ahlinder I, Hansson HB, Melander E, Mölstad S, Söderström M, Persson K. 1997. Duration of nasopharyngeal carriage of penicillin-resistant *Streptococcus pneumoniae*: experiences from the South Swedish Pneumococcal Intervention Project. Clin Infect Dis 25:1113–1117. https://doi.org/10.1086/516103.
48. Sá-Leão R, Nunes S, Brito-Avô A, Alves CR, Carriço JA, Saldanha J, Almeida JS, Santos-Sanches I, de Lencastre H. 2008. High rates of transmission of and colonization by *Streptococcus pneumoniae* and *Haemophilus influenzae* within a day care center revealed in a longitudinal study. J Clin Microbiol 46:225–234. https://doi.org/10.1128/JCM.01551-07.
49. Smith T, Lehmann D, Montgomery J, Gratten M, Riley ID, Alpers MP. 1993. Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children. Epidemiol Infect 111:27–39. https://doi.org/10.1017/S0950268800056648.
50. Sá-Leão R, Pinto F, Aguiar S, Nunes S, Carriço JA, Frazão N, Gonçalves-Sousa N, Melo-Cristino J, de Lencastre H, Ramirez M. 2011. Analysis of invasiveness of pneumococcal serotypes and clones circulating in Portugal before widespread use of conjugate vaccines reveals heterogeneous behavior of clones expressing the same serotype. J Clin Microbiol 49:1369–1375. https://doi.org/10.1128/JCM.01763-10.
51. Foweraker J. 2009. Recent advances in the microbiology of respiratory tract infection in cystic fibrosis. Br Med Bull 89:93–110. https://doi.org/10.1093/bmb/ldn050.
52. del Campo R, Morosini MI, de la Pedrosa EG, Fenoll A, Muñoz-Almagro C, Máiz L, Baquero F, Cantón R, Spanish Pneumococcal Infection Study Network. 2005. Population structure, antimicrobial resistance, and mutation frequencies of *Streptococcus pneumoniae* isolates from cystic fibrosis patients. J Clin Microbiol 43:2207–2214. https://doi.org/10.1128/JCM.43.5.2207-2214.2005.
53. Bauernfeind A, Bertele RM, Harms K, Hörl G, Jungwirth R, Petermüller C, Przyklenk B, Weisslein-Pfister C. 1987. Qualitative and quantitative microbiological analysis of sputa of 102 patients with cystic fibrosis. Infection 15:270–277. https://doi.org/10.1007/BF01644137.
54. May JR, Herrick NC, Thompson D. 1972. Bacterial infection in cystic fibrosis. Arch Dis Child 47:908–913. https://doi.org/10.1136/adc.47.256.908.
55. Esposito S, Colombo C, Tosco A, Montemitro E, Volpi S, Ruggiero L, Lelii M, Bisogno A, Pelucchi C, Principi N, Italian Pneumococcal Study Group on Cystic Fibrosis. 2016. *Streptococcus pneumoniae* oropharyngeal colonization in children and adolescents with cystic fibrosis. J Cyst Fibros 15:366–371. https://doi.org/10.1016/j.jcf.2015.05.008.
56. Thornton CS, Brown EL, Alcantara J, Rabin HR, Parkins MD. 2015. Prevalence and impact of *Streptococcus pneumoniae* in adult cystic fibrosis patients: a retrospective chart review and capsular serotyping study. BMC Pulm Med 15:49. https://doi.org/10.1186/s12890-015-0041-z.
57. Hoiby N, Hoff GE, Jensen K, Lund E. 1976. Serological types of *Diplococcus pneumoniae* isolated from the respiratory tract of children with cystic fibrosis and children with other diseases. Scand J Respir Dis 57:37–40.
58. Pimentel de Araujo F, D'Ambrosio F, Camilli R, Fiscarelli E, Di Bonaventura G, Baldassarri L, Visca P, Pantosti A, Gherardi G. 2014. Characterization of *Streptococcus pneumoniae* clones from paediatric patients with cystic fibrosis. J Med Microbiol 63:1704–1715. https://doi.org/10.1099/jmm.0.072199-0.
59. Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, Bentley SD, García E, Mitchell TJ. 2009. Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. J Bacteriol 191:4854–4862. https://doi.org/10.1128/JB.01272-08.
60. Gladstone RA, Gritzfeld JF, Coupland P, Gordon SB, Bentley SD. 2015. Genetic stability of pneumococcal isolates during 35 days of human experimental carriage. Vaccine 33:3342–3345. https://doi.org/10.1016/j.vaccine.2015.05.021.
61. Avery OT, MacLeod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 79:137–158.

62. Lanie JA, Ng WL, Kazmierczak KM, Andrzejewski TM, Davidsen TM, Wayne KJ, Tettelin H, Glass JI, Winkler ME. 2007. Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. J Bacteriol 189:83–51. https://doi.org/10.1128/JB.01148-06.
63. Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, Janto B, Boissy RJ, Hogg J, Barbadora K, Sampath R, Lonergan S, Post JC, Hu FZ, Ehrlich GD. 2010. Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog 6:e1001108. https://doi.org/10.1371/journal.ppat.1001108.
64. Meyer K, Chaffee E, Hobby GL, Dawson MH. 1941. Hyaluronidases of bacterial and animal origin. J Exp Med 73:309–326. https://doi.org/10.1084/jem.73.3.309.
65. Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, Sørensen UBS. 2008. Evolution of *Streptococcus pneumoniae* and its close commensal relatives. PLoS One 3:e2683. https://doi.org/10.1371/journal.pone.0002683.
66. Li S, Kelly SJ, Lamani E, Ferraroni M, Jedrzejas MJ. 2000. Structural basis of hyaluronan degradation by *Streptococcus pneumoniae* hyaluronate lyase. EMBO J 19:1228–1240. https://doi.org/10.1093/emboj/19.6.1228.
67. Paton JC, Berry AM, Lock RA. 1997. Molecular analysis of putative pneumococcal virulence proteins. Microb Drug Resist 3:1–10. https://doi.org/10.1089/mdr.1997.3.1.
68. Berry AM, Paton JC. 2000. Additive attenuation of virulence of *Streptococcus pneumoniae* by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. Infect Immun 68:133–140. https://doi.org/10.1128/IAI.68.1.133-140.2000.
69. Feldman C, Cockeran R, Jedrzejas MJ, Mitchell TJ, Anderson R. 2007. Hyaluronidase augments pneumolysin-mediated injury to human ciliated epithelium. Int J Infect Dis 11:11–15. https://doi.org/10.1016/j.ijid.2005.09.002.
70. Lacks S, Hotchkiss RD. 1960. A study of the genetic material determining an enzyme activity in Pneumococcus. Biochim Biophys Acta 39:508–518. https://doi.org/10.1016/0006-3002(60)90205-5.
71. Alloing G, Granadel C, Morrison DA, Claverys JP. 1996. Competence pheromone, oligopeptide permease, and induction of competence in *Streptococcus pneumoniae*. Mol Microbiol 21:471–478. https://doi.org/10.1111/j.1365-2958.1996.tb02556.x.
72. Sibold C, Henrichsen J, König A, Martin C, Chalkley L, Hakenbeck R. 1994. Mosaic *pbpX* genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from *pbpX* genes of a penicillin-sensitive *Streptococcus oralis*. Mol Microbiol 12:1013–1023. https://doi.org/10.1111/j.1365-2958.1994.tb01089.x.
73. Hoskins J, Alborn WE, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W, Geringer C, Gilmour R, Glass JS, Khoja H, Kraft AR, Lagace RE, LeBlanc DJ, Lee LN, Lefkowitz EJ, Lu J, Matsushima P, McAhren SM, McHenney M, McLeaster K, Mundy CW, Nicas TI, Norris FH, O'Gara M, Peery RB, Robertson GT, Rockey P, Sun PM, Winkler ME, Yang Y, Young-Bellido M, Zhao G, Zook CA, Baltz RH, Jaskunas SR, Rosteck PR, Skatrud PL, Glass JI. 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 183:5709–5717. https://doi.org/10.1128/JB.183.19.5709-5717.2001.
74. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75.
75. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD. 2011. Rapid pneumococcal evolution in response to clinical interventions. Science 331:430–434. https://doi.org/10.1126/science.1198545.
76. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

2.2 Draft genome sequences of two *Streptococcus pneumoniae* serotype 19A sequence type 226 clinical isolates from Hungary, Hu17 with high-level beta-lactam resistance and Hu15 of a penicillin-sensitive phenotype

Martin Rieger, Dalia Denapaite, Reinhold Brückner, Patrick Maurer and Regine Hakenbeck. Genome Announc. 2017 May; 5(20): e00401-17.

Summary:

Annotated draft genomes of two members (Hu15 and Hu17) of the high-level penicillin resistant *Streptococcus pneumoniae* Hu19A-6 clone ST226 were generated. This clone appears more variable compared to other common *S. pneumoniae* clones judged from previous genome hybridization data using an oligonucleotide microarray of a representative *S. pneumoniae* TIGR4 strain (Tettelin, et al., 2001), and this finding can now be analysed in detail on the genome sequence level. Moreover, the two strains represent a unique situation since strain Hu15 is penicillin sensitive, whereas Hu17 is highly resistant against beta-lactam antibiotics similar to other members of this clone. Thus, the genome sequences of these two strains offer the opportunity of further investigations on the evolution of penicillin resistance.

Own contribution to the paper:

Assembly of sequence reads and final generation of genome sequences from assembled contigs including annotation of two *S. pneumoniae* ST226 isolates (Hu15 and Hu17) and submission of genome sequences to the NCBI database. Comparative analysis of the two genomes and their plasmids including single nucleotide variation (SNV) retrieval and analysis of diverging sequence regions. Manual SNV retrieval and confirmation of results of automatically performed analysis by a newly developed wrapper tool. Comparison of the two ST226 genomes with the genome sequence of *S. pneumoniae* Hungary^{19A}-6 (Acc. No. NC_010380; afterwards referred to as Hu19A). Extraction of coding regions and deduction of proteins for comparison with *S. pneumoniae* Hu19A. Unpublished work is described in chapter 3.2

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

Draft Genome Sequences of Two *Streptococcus pneumoniae* Serotype 19A Sequence Type 226 Clinical Isolates from Hungary, Hu17 with High-Level Beta-Lactam Resistance and Hu15 of a Penicillin-Sensitive Phenotype

Martin Rieger,* Dalia Denapaite, Reinhold Brückner, Patrick Maurer,* Regine Hakenbeck

Department of Microbiology, University of Kaiserslautern, Kaiserslautern, Germany

ABSTRACT The draft genome sequences of two multiple-antibiotic-resistant *Streptococcus pneumoniae* isolates from Hungary, Hu15 and Hu17, are reported here. Strain Hu15 is penicillin susceptible, whereas Hu17 is a high-level-penicillin-resistant strain. Both isolates belong to the serotype 19A sequence type 226, a single-locus variant (in the *ddl* locus) of the Hungary^{19A}-6 clone.

High-level-penicillin- and multiple-antibiotic-resistant *Streptococcus pneumoniae* (PRSP) strains of serotype 19A were prevalent in Hungary during the 1990s (1, 2). The strain HUN663 represents the clone Hungary^{19A}-6, as defined by multilocus sequence typing (MLST), belonging to sequence type 268 (ST268) (3). Meanwhile, another 11 strains representing 4 single-locus variants (SLVs) of Hungary^{19A}-6 with respect to the *ddl* locus are reported in the MLST database (http://pubmlst.org/spneumoniae/), including isolates from the Czech Republic and Slovakia (4) and Norway belonging to ST226, ST340, ST382, and ST7133. The *ddl* locus maps closely to *pbp2b*, which is acquired by horizontal gene transfer during the evolution of penicillin resistance (5) and is therefore not used in phylogenetic analyses based on MLST. Therefore, all these strains can be considered to belong to the clone Hungary^{19A}-6. In agreement, strains of ST258 and ST226 express a penicillin-binding protein 3 (PBP3) of different electrophoretic mobility compared to that of most other *S. pneumoniae* (1, 6, 7) strains.

Isolates of ST226 are varied in their MIC values, PBP profile, and PBP2x sequences, and multilocus electrophoretic typing revealed several electrophoretic types (6, 8). Accordingly, their genomes appear surprisingly variable compared to other clones (9); in fact, Hungary^{19A}-6 had acquired the largest proportion of genes (8.2%) from *Streptococcus mitis* in a comparative genomic analysis of 35 *Streptococcus* species genomes (10). These strains are part of the Kaiserslautern strain collection obtained from Anna Marton as cited in (6) now held at the German National Reference Center for Streptococci in Aachen, Germany. Interestingly, one member (strain Hu15) was penicillin susceptible (MICs, 0.1 μg/ml for oxacillin and 0.024 μg/ml for cefotaxime), whereas Hu17 is among the strains expressing the highest beta-lactam resistance levels (MICs, 30 and 1.6 μg/ml, respectively). Moreover, the strains were resistant to tetracycline, streptomycin, erythromycin, and trimetroprim.

Information retrieved from these two genomes will help decipher the evolutionary pathway of penicillin resistance. The genome sequences revealed a plasmid related to pSpnP1 (11).

Received 3 April 2017 **Accepted** 4 April 2017 **Published** 18 May 2017

Citation Rieger M, Denapaite D, Brückner R, Maurer P, Hakenbeck R. 2017. Draft genome sequences of two *Streptococcus pneumoniae* serotype 19A sequence type 226 clinical isolates from Hungary, Hu17 with high-level beta-lactam resistance and Hu15 of a penicillin-sensitive phenotype. Genome Announc 5:e00401-17. https://doi.org/10.1128/genomeA.00401-17.

Address correspondence to Regine Hakenbeck, hakenb@rhrk.uni-kl.de.

* Present address: Martin Rieger, neox Aktiengesellschaft für Informationstechnologie, Pirmasens, Germany; Patrick Maurer, Hochschule für Technik und Wirtschaft des Saarlandes, Saarbrücken, Germany.

The genomes were sequenced using an Illumina HiSeq platform (Hu15/Hu17, 2,203,997/2,192,370 bp of paired-end reads). Genomes were assembled using gsAssembler (version 2.6). RATT was used for genome annotation (12) using *S. pneumoniae* Hu19A-6 as a reference genome, and was adjusted manually if necessary, according to NCBI submission guidelines. The five capped small RNA (csRNA) genes encoding small regulatory RNAs controlled by CiaRH (13) were identified by their high identity to the *S. pneumoniae* R6 counterparts and added to the annotation. The genomes of Hu15/Hu17 were assembled into 176/200 contigs, with a total length of 2,136,165/2,141,026 nucleotides (nt) (sequenced to ~2.2 million reads with ~157/159× coverage). A plasmid was assigned to 1/4 contig(s), with a total length of 5,327/5,112 nt. The predicted genes from the genomes include 2,191/2,190 coding sequences (CDSs), 64/93 incomplete genes at contig ends, 2/3 rRNAs, 22/23 tRNAs, 5/4 csRNAs, and 3/3 other small RNAs (RNase P; *srpB*; transfer-messenger RNA [tmRNA]).

Accession number(s). The draft genome sequences and plasmid sequences of Hu15 and Hu17 have been deposited in the NCBI database under the GenBank accession numbers CP020551 and CP020552 and CP020549 and CP020550, respectively.

ACKNOWLEDGMENT

This work was supported by a grant from the Deutsche Forschungsgemeinschaft DFG Ha1011/11-3 to R.H.

REFERENCES

1. Marton A, Gulyas M, Muñóz R, Tomasz A. 1991. Extremely high incidence of antibiotic resistance in clinical isolates of *Streptococcus pneumoniae* in Hungary. J Infect Dis 163:542–548. https://doi.org/10.1093/infdis/163.3.542.
2. Marton A, Mészner Z. 1999. Epidemiological studies on drug resistance in *Streptococcus pneumoniae* in Hungary: an update for the 1990s. Microb Drug Resist 5:201–205. https://doi.org/10.1089/mdr.1999.5.201.
3. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, Hakenbeck R, Hryniewicz W, Lefévre JC, Tomasz A, Klugman KP. 2001. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the Pneumococcal Molecular Epidemiological Network (PMEN). J Clin Microbiol 39:2565–2571. https://doi.org/10.1128/JCM.39.7.2565-2571.2001.
4. Figueiredo AM, Austrian R, Urbaskova P, Teixeira LA, Tomasz A. 1995. Novel penicillin-resistant clones of Streptococcus pneumoniae in the Czech Republic and in Slovakia. Microb Drug Resist 1:71–78. https://doi.org/10.1089/mdr.1995.1.71.
5. Enright MC, Spratt BG. 1999. Extensive variation in the *ddl* gene of penicillin-resistant *Streptococcus pneumoniae* results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. Mol Biol Evol 16:1687–1695. https://doi.org/10.1093/oxfordjournals.molbev.a026082.
6. Reichmann P, Varon E, Günther E, Reinert RR, Lüttiken R, Marton A, Geslin P, Wagner J, Hakenbeck R. 1995. Penicillin-resistant *Streptococcus pneumoniae* in Germany: genetic relationship to clones from other European countries. J Med Microbiol 43:377–385. https://doi.org/10.1099/00222615-43-5-377.
7. Krauss J, Hakenbeck R. 1997. A mutation in the D,D-carboxypeptidase penicillin-binding protein 3 of *Streptococcus pneumoniae* contributes to cefotaxime resistance of the laboratory mutant C604. Antimicrob Agents Chemother 41:936–942.
8. Reichmann P, König A, Marton A, Hakenbeck R. 1996. Penicillin-binding proteins as resistance determinants in clinical isolates of *Streptococcus pneumoniae*. Microb Drug Resist 2:177–181. https://doi.org/10.1089/mdr.1996.2.177.
9. Hakenbeck R, Balmelle N, Weber B, Gardès C, Keck W, de Saizieu A. 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies variation of *Streptococcus pneumoniae*. Infect Immun 69:2477–2486. https://doi.org/10.1128/IAI.69.4.2477-2486.2001.
10. Kilian M, Riley DR, Jensen A, Brüggemann H, Tettelin H. 2014. Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. mBio 5:e01490-14. https://doi.org/10.1128/mBio.01490-14.
11. Romero P, Llull D, García E, Mitchell TJ, López R, Moscoso M. 2007. Isolation and characterization of a new plasmid pSpnP1 from a multidrug-resistant clone of *Streptococcus pneumoniae*. Plasmid 58:51–60. https://doi.org/10.1016/j.plasmid.2006.12.006.
12. Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res 39:e57. https://doi.org/10.1093/nar/gkq1268.
13. Halfmann A, Kovács M, Hakenbeck R, Brückner R. 2007. Identification of the genes directly controlled by the response regulator CiaR in *Streptococcus pneumoniae*: five out of fifteen promoters drive expression of small noncoding RNAs. Mol Microbiol 66:110–126. https://doi.org/10.1111/j.1365-2958.2007.05900.x.

2.3 Highly variable *Streptococcus oralis* strains are common among viridans *Streptococci* isolated from primates

Dalia Denapaite, **Martin Rieger**, Sophie Köndgen, Reinhold Brückner, Irma Ochigava, Peter Kappeler, Kerstin Mätz-Rensing, Fabian Leendertz and Regine Hakenbeck. mSphere. 2016 Mar-Apr; 1(2): e00041-15.

Summary:

Viridans streptococci represent a major part of the commensal flora of the human upper respiratory tract. *Streptococcus pneumoniae* is the only member of this group of bacteria which shows a distinct virulence potential. The pathogenicity is dependent on a set of virulence factors including the polysaccharide capsule and a variety of surface proteins such as choline binding proteins, the hyaluronidase HlyA, and the cytolysin pneumolysin. Many other factors described as virulence factors are also present in non-pathogen members of the viridans group and are most likely important for bacterium/host interaction. Based on comparative genetic analysis it has been proposed that *S. pneumoniae* and its close relatives *Streptococcus mitis* and *S. pseudopneumoniae* share a common ancestor and have evolved later compared to *S. oralis* (Kilian, et al., 2008). *S. pneumoniae* is considered to be a human specific pathogen, but it is not known whether this is true for the closely related species as well. The current study determined the distribution of viridans streptococci isolated from great apes and other monkeys (captivity and free living) to characterize the distribution of streptococcal species among primates. Moreover, a detailed comparative analysis of genome sequences obtained from different streptococcal species was performed, focusing on the presence of pneumococcal virulence factors, large pneumococcal genomic islands, small non-coding RNA controlled by CiaRH (csRNA), and genes involved in the synthesis of the important surface polymers peptidoglycan and teichoic acids.

This study revealed that *S. oralis* could only be found in Old World monkeys, providing evidence that this species evolved prior to the origin of human. In addition, *S. oralis* was also isolated from Rhesus monkeys held in captivity; further investigations will be necessary to confirm its presence in wild animals, since transfer from human to monkeys cannot be excluded at this stage.

The genomic analysis revealed that many pneumococcal virulence factors are present in the streptococcal genomes analysed here. Only a few genes encoding surface proteins appear to be present in *S. pneumoniae* and only rarely or not at all in other streptococcal species. The pneumococcal neuraminidases NanBC occurred only in one *S. mitis* strain, but a related protein, a putative β-N-acetyl-hexosaminidase occurred in all *S. oralis* strains.

Genes encoding the penicillin-binding proteins PBP2x, PBP2b and PBP1a and MurMN also involved in peptidoglycan synthesis were surprisingly variable in *S. oralis*. Interestingly, some species contained two homologs of PBP3, named group 1 (the common PBP3) and group 2. In these genomes, the group 2 PBP3 was intact, while the group 1 gene appeared inactive judged from the presence or absence of the conserved active site motifs. Surprisingly, besides an unusual MurM, MurN is absent in most *S. oralis* genomes. At least three variants of the genes responsible for choline decoration of teichoic acids (*lic* clusters) were identified among *S. oralis/S. mitis*, suggesting also a distinct biochemistry of these surface polymers.

This study has revealed some important features in streptococci that will help to unravel important questions such as their adaptation to diverse habitats and mechanisms involved in diversification of their genomes.

Own contribution to the paper:

Assembly of sequence reads and final generation of scaffold sequences from assembled contigs including annotation of 30 bacterial isolates (DD01-DD30) and submission of scaffold sequences to the NCBI database. Comparative analysis of proteins and pilus cluster common to *S. oralis* are described in chapter 3.3.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere™

RESEARCH ARTICLE
Ecological and Evolutionary Science

Highly Variable *Streptococcus oralis* Strains Are Common among Viridans Streptococci Isolated from Primates

Dalia Denapaite,[a] Martin Rieger,[a] Sophie Köndgen,[b] Reinhold Brückner,[a] Irma Ochigava,[a] Peter Kappeler,[c] Kerstin Mätz-Rensing,[c] Fabian Leendertz,[b] Regine Hakenbeck[a]

Department of Microbiology, University of Kaiserslautern, Kaiserslautern, Germany[a]; Project Group 3 Epidemiology of Highly Pathogenic Microorganisms, Robert Koch-Institute, Berlin, Germany[b]; Behavioral Ecology and Sociobiology Unit, German Primate Center, Göttingen, Germany[c]

ABSTRACT Viridans streptococci were obtained from primates (great apes, rhesus monkeys, and ring-tailed lemurs) held in captivity, as well as from free-living animals (chimpanzees and lemurs) for whom contact with humans is highly restricted. Isolates represented a variety of viridans streptococci, including unknown species. *Streptococcus oralis* was frequently isolated from samples from great apes. Genotypic methods revealed that most of the strains clustered on separate lineages outside the main cluster of human *S. oralis* strains. This suggests that *S. oralis* is part of the commensal flora in higher primates and evolved prior to humans. Many genes described as virulence factors in *Streptococcus pneumoniae* were present also in other viridans streptococcal genomes. Unlike in *S. pneumoniae*, clustered regularly interspaced short palindromic repeat (CRISPR)–CRISPR-associated protein (Cas) gene clusters were common among viridans streptococci, and many *S. oralis* strains were type PI-2 (pilus islet 2) variants. *S. oralis* displayed a remarkable diversity of genes involved in the biosynthesis of peptidoglycan (penicillin-binding proteins and MurMN) and choline-containing teichoic acid. The small noncoding *cia*-dependent small RNAs (csRNAs) controlled by the response regulator CiaR might contribute to the genomic diversity, since we observed novel genomic islands between duplicated csRNAs, variably present in some isolates. All *S. oralis* genomes contained a *β*-*N*-acetyl-hexosaminidase gene absent in *S. pneumoniae*, which in contrast frequently harbors the neuraminidases NanB/C, which are absent in *S. oralis*. The identification of *S. oralis*-specific genes will help us to understand their adaptation to diverse habitats.

IMPORTANCE *Streptococcus pneumoniae* is a rare example of a human-pathogenic bacterium among viridans streptococci, which consist of commensal symbionts, such as the close relatives *Streptococcus mitis* and *S. oralis*. We have shown that *S. oralis* can frequently be isolated from primates and a variety of other viridans streptococci as well. Genes and genomic islands which are known pneumococcal virulence factors are present in *S. oralis* and *S. mitis*, documenting the widespread occurrence of these compounds, which encode surface and secreted proteins. The frequent occurrence of CRISP-Cas gene clusters and a surprising variation of a set of small noncoding RNAs are factors to be considered in future research to further our understanding of mechanisms involved in the genomic diversity driven by horizontal gene transfer among viridans streptococci.

KEYWORDS: *Streptococcus oralis*, horizontal gene transfer, primates, teichoic acid, viridans streptococci, virulence factors

Received 29 October 2015 **Accepted** 6 February 2016 **Published** 9 March 2016

Citation Denapaite D, Rieger M, Köndgen S, Brückner R, Ochigava I, Kappeler P, Mätz-Rensing K, Leendertz F, Hakenbeck R. 2016. Highly variable *Streptococcus oralis* strains are common among viridans streptococci isolated from primates. mSphere 1(2):e00041-15. doi: 10.1128/mSphere.00041-15.

Editor Melanie Blokesch, Swiss Federal Institute of Technology, Lausanne, Switzerland

Address correspondence to Regine Hakenbeck, hakenb@rhrk.uni-kl.de.

Viridans streptococci are a major part of the commensal microbiota of the upper respiratory tract of humans (for a review, see reference 1). Only one member, *Streptococcus pneumoniae*, represents a major pathogen causing a variety of diseases, such as pneumonia, otitis media, sepsis, and meningitis, whereas even its closest relatives *Streptococcus mitis*, *Streptococcus oralis*, and the recently described *Streptococcus pseudopneumoniae* (2) are rarely associated with disease. Genomic analyses represent a powerful tool to further our understanding of the genetic basis for the pathogenic potential of *S. pneumoniae* and to decipher the evolutionary relationship between these species.

Many gene products have been described as virulence factors in *S. pneumoniae*. Strains carrying mutations in these genes are less virulent in mouse models. However, most of them are probably required for host interaction rather than being directly responsible for disease since in most cases homologs are present in *S. pneumoniae*'s nonpathogenic commensal relatives *S. mitis* and *S. pseudopneumoniae* (2–4). There are only a few components that are clearly associated with *S. pneumoniae* and which are not or are only rarely found in commensal streptococci: the pneumolysin Ply, a pore-forming toxin whose gene is located on an islet together with the autolysin gene *lytA*, and the three choline-binding proteins (CBPs) PspA, PcpA, and PspC, including the Hic variant of Psp and the hyaluronidase HlyA. Moreover, the biochemically highly diverse capsule is essential for pneumococcal pathogenicity. Nevertheless, even those compounds are either absent (rare) in *S. pneumoniae* or highly varied. A random distribution of virulence genes (*lytA*, *ply*, and the *cap* locus, representing the capsule biosynthesis operon) has been observed among *S. mitis* strains (5), suggesting that the acquisition and loss of genes is an ongoing process in this group of bacteria. It has been proposed that the three species *S. mitis*, *S. pneumoniae*, and *S. pseudopneumoniae* arose from an ancient bacterial population that included all *S. pneumoniae*-specific genes (6), and genomic analysis of 35 *Streptococcus* spp. indicated that the common ancestor was a pneumococcus-like species (5). Moreover, the authors provided evidence that interspecies gene transfer occurred mainly unidirectionally from *S. mitis* to *S. pneumoniae*.

Due to the ability of streptococci to develop genetic competence resulting in a large accessory genome, identification to the species level has been problematic using phenotypic and physiological criteria. Therefore, a variety of genotypic methods have been established based on sequences from the core genes (housekeeping genes). Multilocus sequence typing (MLST) has become the gold standard for clonal analysis, especially within species (7), and multilocus sequence analysis (MLSA) has been applied to differentiate closely related streptococcal species (8). Nonetheless, recombinogenic bacteria do not form clusters with clear boundaries and have been termed "fuzzy species" (9). Phylogenetic analyses show that *S. pneumoniae* is a single lineage in a cluster formed by a variety of *S. mitis* clusters, whereas the *S. oralis* cluster is well separated (5, 6, 10). Diversification within the *S. pneumoniae* lineage has probably occurred during growth of the human population, its primary host (5). This suggests that the diversification of *S. mitis* and *S. oralis* took place earlier and that great apes might still harbor *S. mitis* and *S. oralis* as part of their commensal flora.

We therefore investigated the distribution of viridans streptococci among great apes and other primates. Samples were obtained from animals held in captivity, namely, gorillas, orangs, and bonobos from the Frankfurt Zoo and rhesus monkeys and ring-tailed lemurs (*Lemur catta*) from the German Primate Center, as well as from free-living animals for whom contact with humans is highly restricted, namely, chimpanzees from the Taï National Park, Ivory Coast, lemurs from the Kirindy Forest in Madagascar (Verreaux's sifaka, *Propithecus verreauxi*), red-fronted lemurs (*Eulemur rufifrons*), Western fat-tailed dwarf lemurs (*Cheirogaleus medius*), and gray mouse lemurs (*Microcebus murinus*). In the first part of our study, the species are described based on genotypic methods. *S. oralis* was common among great apes, including wild chimpanzees, and was also found in rhesus monkeys. In the second part of our study, the genomes from 23 isolates are analyzed in detail. Emphasis is placed on the presence of large genomic islands which are part of the accessory genome of *S. pneumoniae*, small noncoding

RNAs controlled by a highly conserved two-component system, CiaRH, *S. pneumoniae* virulence factors, and genes involved in peptidoglycan and teichoic acid (TA) metabolism.

RESULTS

Determination to the species level of viridans isolates from primates by MLSA. All isolates from primates were initially characterized for their morphologies (by colony and microscopic analysis of cells), antibiotic susceptibilities, 16S rRNAs, and in some cases SDS-PAGE protein patterns of cell lysates and penicillin binding protein (PBP) profiles. A total of 139 isolates was included in further analyses (see Table S1 in the supplemental material). MICs of oxacillin above 1 μg/ml and resistance to antibiotics of other classes were common among isolates from animals held in captivity, but isolates from free-living chimpanzees of the Taï National Park in Africa did not show significant resistance patterns for any of the antibiotics tested (Table S1). If a group of isolates obtained from one animal showed properties identical to those listed above, only one isolate was subjected to further analysis by MLST. This group consisted of 44 isolates, including all isolates from Taï chimpanzees and at least three isolates from the other group of animals. Most isolates from zoo animals were furthermore defined by MLST analysis (see below).

Figure 1 shows a neighbor-joining radial tree of the primate isolates in comparison with the published MLSA data set of a wide variety of streptococcal species (8). In addition, MLSA data were extracted from another five genomes available recently. *Streptococcus oligofermentans* AS1.3089 (11) was positioned within the Anginosus group, as expected, whereas the reference strain used by Bishop et al. (8), *S. oligofermentans* CCUG48365, clustered within *S. oralis*, strongly suggesting that the latter strain has been misidentified as suggested before (8). *Streptococcus sinensis* HKU4 (12) was located at nearly the same position as *S. sinensis* SK1972 (8), and *Streptococcus downei* F0415 was part of the Mutans group. Moreover, *Streptococcus tigurinus* AZ_3a and *Streptococcus dentisani* 7747, representing two new species that have been described as being closely related to *S. oralis* (13–15), were located within the *S. oralis* group (Fig. 1) and will be discussed below. The primate isolates covered a wide range of *Streptococcus* spp. Sixteen isolates obtained from zoo animals, wild chimpanzees, and rhesus monkeys were found among the *S. oralis* cluster. One isolate from a gorilla (DD22) clustered among *S. mitis* strains, and one from another gorilla (DD18) clustered close to *Streptococcus infantis*/*Streptococcus peroris*. Three isolates from Taï chimpanzees were defined as *Streptococcus gordonii* (CB10, CB18, and DD07), DD08 was defined as *Streptococcus cristatus*, and DD04 from ring-tailed lemurs mapped close to *S. sinensis*, all of which are within the Mitis group of viridans streptococci. One isolate (DD09) from a chimpanzee was defined as *Streptococcus constellatus*, which belongs to the Anginosus group of viridans streptococci. No streptococci could be isolated from free-living lemurs of Madagascar, where *Enterococcus* sp. and *Lactococcus* sp. were commonly obtained. Only from *Propithecus verreauxii* were we able to obtain *Streptococcus* sp. that could not be defined by MLSA (Table S1).

A total of 20 isolates fell into different lineages outside the streptococcal species included in the MLSA tree (Fig. 1). A cluster of seven isolates from Madagascar lemurs is magnified in Fig. 1B. Some isolates could not be identified to the species level by MLSA. From Taï chimpanzees, there were two clusters within the Mitis group consisting of three strains (CB23, CB11, and DD10) and four strains (DD11, CB9, CB17, and CB20) and two pairs of strains (DD12/CB5 and DD13/CB14) between the Anginosus and Salivarius groups that could not be identified to the species levels; from ring-tailed lemurs, there were two isolates between the Salivarius group and *S. pyogenes* (KG3e and DD06) that could not be identified to the species level. Comparison of their 16S rRNA sequences did not match any known sequences from the NCBI data bank except those of DD06 (*Streptococcus lutetiensis*) and the Madagascar lemurs (*Streptococcus gallolyticus*) (Fig. 1B). It should be noted that 16S rRNA sequences pose significant

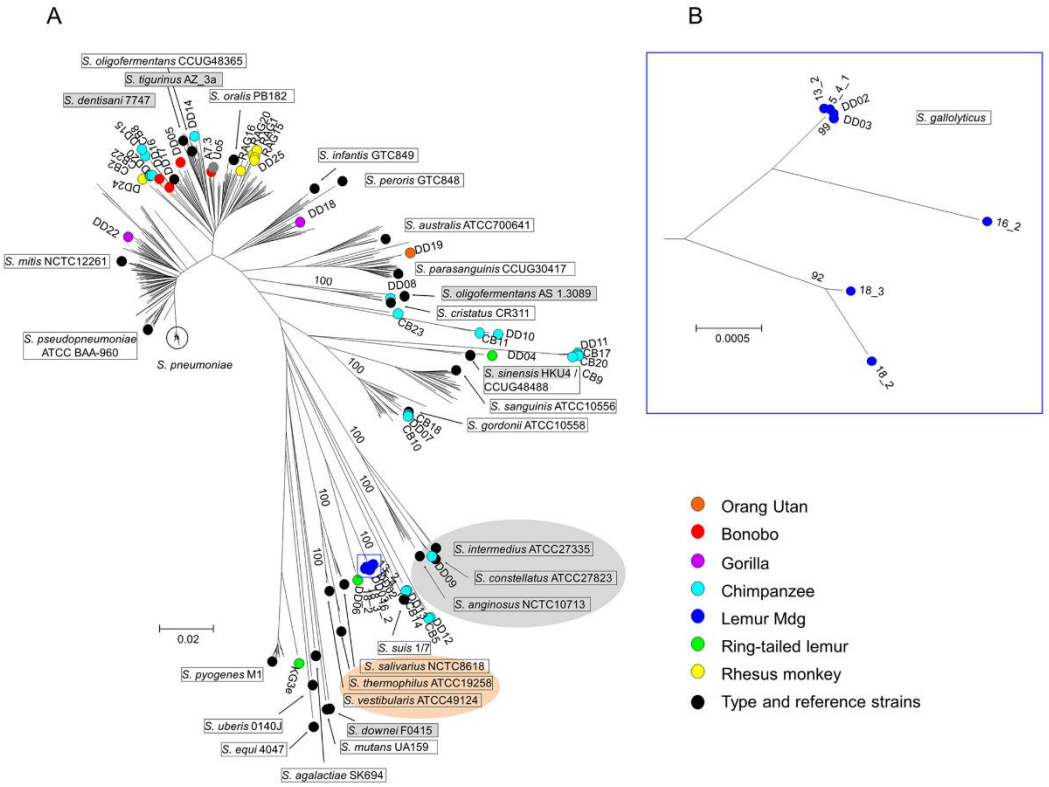


FIG 1 MLSA trees of strains from this study and reference strains. (A) A neighbor-joining tree was constructed using the concatenated sequences of the MLSA loci from 44 strains of this study combined with sequences of 427 strains from the study of Bishop et al. (8). In addition, MLSA genes were extracted from the genomes of *S. tigurinus* AZ_3a (GCF_000344275.1), *S. dentisani* 7747 (GCF_000382805.1), *S. sinensis* HKU4 (GCF_000767835.1), *S. oligofermentans* AS1.3089 (CP004409.1), and *S. downei* F0415 (GCA_000180055.1) (strains included in addition to those from the study by Bishop et al. [8] are shaded gray). Viridans group reference and type strains are framed. The Anginosus group of viridans streptococci is shaded gray and the Salivarius group orange. The color key for reference strains and isolates from primates is indicated on the right. Mdg, Madagascar. (B) Neighbor-joining tree of the *Streptococcus* sp. cluster of strains from Madagascar lemurs (blue circles in the square in panel A) magnified to show more clearly the clustering of the strains. Bootstrap values (percentages) are based on 1,000 replications. The bar refers to genetic divergence as calculated by the MEGA software.

problems for identification, and matches below 100% are not very meaningful to differentiate between species (see reference 16 and references therein).

Determination to the species level by genome analysis and plasmids. Twenty-five streptococcal isolates (named "DD" followed by consecutive numbers) were chosen for whole-genome sequencing, 23 of which were *Streptococcus* spp. (see Table S1 in the supplemental material). This included 8 *S. oralis* isolates representing different lineages of the MLSA tree to cover a broad range of variation within this species. In addition, we used 1 isolate each of *S. gordonii*, *S. cristatus*, *S. constellatus*, *S. infantis*, and *S. mitis* as defined by MLSA and 10 isolates of unclear species determination according to MLSA.

The species defined by MLSA were confirmed by genome sequences (Table S2). According to BLAST analysis with 16S rRNA, MLSA genes, and *pbp2a* in the NCBI microbial genome data bank (Table S3), the two genomes of DD02 and DD03 from Lemur isolates were identified as *S. gallolyticus*, DD06 from ring-tailed lemurs was identified as *S. lutetiensis*, and DD19 from zoo animals was identified as *S. parasanguinis*. DD04 was close to *S. sinensis*. There remained four isolates from Taï chimpanzees

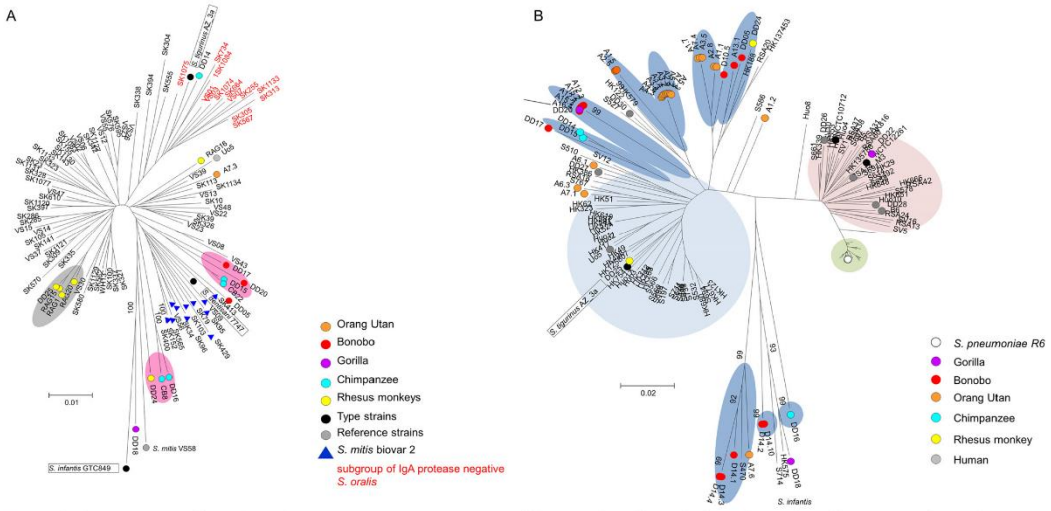


FIG 2 Phylogenetic trees of *S. oralis* and *S. mitis* strains. (A) Tree generated by MLSA loci of *S. oralis* from this study and sequences of *S. oralis* strains, *S. tigurinus* AZ_3a, and *S. dentisani* 7747 from the MLSA tree shown in Fig. 1. Gray shading, cluster of isolates from rhesus monkeys; pink shading, lineages consisting of primate isolates only. One *S. mitis* strain from the study of Bishop et al. (8), VS58, was included for comparison. Red letters indicate the subgroup of IgA protease-negative *S. oralis* isolates specified in reference 8. Bootstrap values (percentages) are based on 1,000 replications. (B) Phylogenetic relationship of primate isolates, including 38 isolates from zoo animals determined by MLST (7) but excluding *ddl*, combined with sequences of 119 human isolates (*S. oralis*, *S. mitis*, and *S. pneumoniae*) from different geographic locations (10) (isolate numbers are preceded by Hu for Hungary, RSA for South Africa, S for Spain, HK for Hong Kong, and B for Germany). Lineages within the *S. oralis* cluster containing only isolates from primates are shaded in dark blue. Bootstrap values (percentages) are based on 500 replications. The bar refers to genetic divergence as calculated by the MEGA software.

(DD10, DD11, DD12, and DD13) whose species could not be determined; for none of their genes did we find close matches in the NCBI data bank.

During preparation of chromosomal DNA, plasmids were detected in five samples. The plasmid of *S. oralis* strain DD25 from rhesus monkeys was identical to the *S. pneumoniae* pSpnP1 plasmid, large parts of which were also found in *S. oralis* strain DD24. Fragments related to pSpnP1 were also present in plasmids from *S. oralis* strain DD17, and *S. infantis* strain DD18 was partially related to a plasmid from *S. pseudopneumoniae* IS7493 pDRPIS7493. No significant matches to the *S. gallolyticus* strain DD03 plasmid were found by BLAST analysis.

A closer look at *S. oralis*. The published MLSA data set distinguishes three phenotypically distinct subclusters among *S. oralis* strains: one which covered strains previously defined as *S. mitis* biovar 2, an IgA protease-negative *S. oralis* cluster, and various lineages of IgA protease-positive *S. oralis* (8). As can be seen in Fig. 2A, S was found within the *S. oralis* subcluster of IgA protease-negative strains and *S. dentisani* among the subcluster of strains previously defined as *S. mitis* biovar 2 (8). We found only one isolate from primates within the biovar 2 group (DD05 from a bonobo) and one within the IgA protease-negative group (DD14) (Fig. 2A). Four isolates from rhesus monkeys formed one subcluster (grey in Fig. 2A). Seven isolates were found on different lineages outside the main *S. oralis* group of human *S. oralis* strains (pink in Fig. 2A).

We then analyzed 38 isolates from zoo primates by MLST since most of them were suspected of being *S. oralis* based on the preliminary characterization. The MLST sequences extracted from the genomes of isolates from rhesus monkeys and chimpanzees were included. The results were compared to previously published MLST data (10) derived from a set of 119 *S. pneumoniae*, *S. mitis*, and *S. oralis* isolates from different geographic areas, and the MLST sequences from *S. tigurinus* AZ_4a were also included

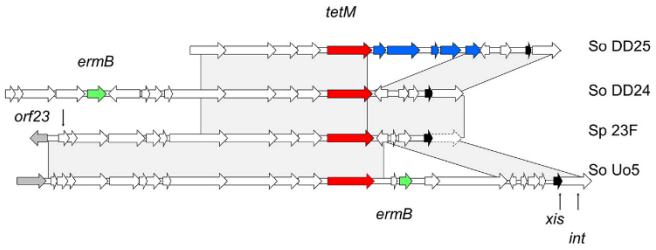


FIG 3 Comparison of TetM-containing genetic elements in two genomes from rhesus monkeys with those of *S. pneumoniae* Spain 23F-1 (Sp 23F) and *S. oralis* Uo5 (So Uo5). Red, *tetM*; green, *ermB*; blue, homology to *Enterococcus faecium* plasmid pM7M2 genes. *xis* is the excisionase gene (black); the integrase is a pseudogene in *S. pneumoniae* (dashed arrow). Gray areas indicate BLASTn matches between sequences.

(Fig. 2B) (not all MLST sequences from the genome of *S. dentisani* 7747 were included since the genes *spi* and *gdh* were not found in the genome). The presence of multiple subclusters within the *S. oralis* cluster is evident also in the MLST-based phylogeny shown here, which positions *S. tigurinus* within the main *S. oralis* cluster. Again, most *S. oralis* isolates from primates except four (A7.1, A6.3, and A6.1 from zoo apes and DD25 from a rhesus monkey) were located outside the main cluster of human *S. oralis* strains in several lineages (blue in Fig. 2B). In only one case were identical MLST sequences obtained from strains from two different primate species: DD20 from a gorilla, A12.3 from bonobo Ku, and A15.2/A15.3 from bonobo Zo (arrow in Fig. 2B; see Table S1 in the supplemental material). DD18, which was defined by MLSA as *S. infantis*, and DD22, identified as *S. mitis*, were positioned in the tree outside the *S. oralis* lineages.

Thus, MLSA as well as MLST data showed that most *S. oralis* isolates from primates—independently of whether they had been obtained from animals held in captivity or from free-living animals—were distinct from those of the many human isolates from different geographic areas, including China, South Africa, and several eastern and western European countries.

Dissemination of large genomic islands. In the following analyses, we included another three human *S. oralis* and four *S. mitis* isolates described before (10, 17). Two isolates from rhesus monkeys (*S. oralis* strains DD24 and DD25) were tetracycline resistant; DD24 was erythromycin resistant as well. Tetracycline resistance in *S. pneumoniae* is most commonly conferred by *tetM*, located on an integrative conjugative element (ICE) of the Tn*916* family (18, 19). As shown in Fig. 3, both genomes carried parts of Tn*916* like the *S. pneumoniae* Spain 23F-1 clone (20). *S. oralis* strain DD24 contained *ermB* located on an ICE which is also present in a variety of *Streptococcus* sp. genomes, including that of *S. pneumoniae* Hungary 19A-6. In contrast, the human isolate *S. oralis* Uo5 contained *ermB* next to *tetM*, a genotype frequently also found in *S. pneumoniae* (18). The *S. oralis* strain DD25 Tn*916* region had an insert corresponding to *Enterococcus faecium* plasmid pM7M2 sequences (21) which are present in a wide variety of Gram-positive bacteria, including *Staphylococcus* spp., *Bacillus* spp., and *Streptococcus* spp., according to BLAST analysis with the NCBI nucleotide data bank (blue in Fig. 3).

There are several large gene clusters of the accessory genome in *S. pneumoniae* (>10 kb) implicated in modulation of the pathogenicity potential (22) and which are found to be widespread among different species. One cluster harbors genes encoding a serine-rich cell surface protein (named PsrP in *S. pneumoniae* and MonX in *S. mitis* B6) with accessory components responsible for glycosylation and export. Serine-rich proteins are adhesins common among Gram-positive bacteria and contribute to a variety of diseases (for a review, see reference 23). This cluster was widespread also among the

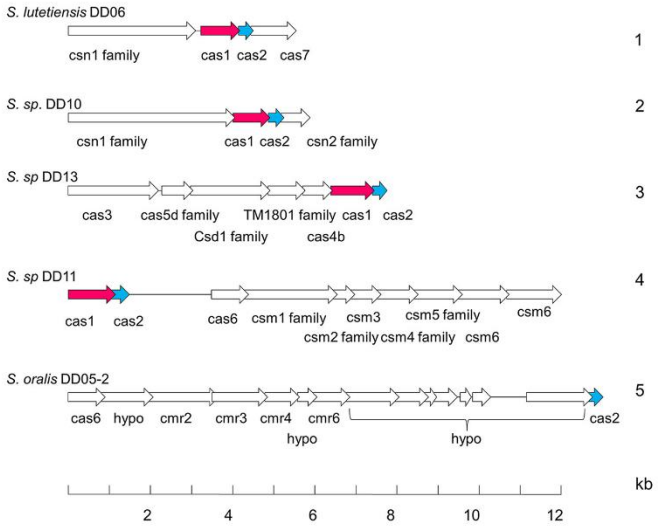


FIG 4 CRISPR-Cas gene clusters in streptococcal genomes. The five classes (designated 1 to 5) of CRISPR-Cas gene clusters identified in the genomes of this study are shown; representative genomes containing these clusters are indicated on the left. The annotation given by the RAST server (http://rast.nmpdr.org/) was used. Red, Cas1 genes; blue, Cas2 genes; *S. sp*, Streptococcus sp.

primate genomes (Table S2). Moreover, a region containing genes for a V-type ATPase was present in several primate genomes (Table S2).

CRISPR-Cas (*c*lustered *r*egularly *i*nterspaced *s*hort *p*alindromic *r*epeats–CRISPR-*as*sociated proteins) loci represent defense systems against foreign genetic elements. Although *S. pneumoniae* does not contain CRISPR sequences, they were found among *S. mitis* and *S. oralis* isolates (5), but information concerning other streptococcal species is still limited (24). We detected CRISPR-Cas gene clusters in most of the streptococcal genomes, and several genomes contained more than one CRISPR-Cas cluster at different genomic positions (Table S2). Five different cluster arrangements were observed (examples are shown in Fig. 4); *S. oralis* genomes contained clusters of types 1, 2, and 5. Four of these clusters contain Cas1 genes which clustered according to their genomic arrangement (Fig. S1), as described by Makarova et al. (24).

An interesting case of intra- and interspecies gene transfer events among the Mitis group of streptococci is the presence of new variants of pilus islet 2 (PI-2), described to occur in *S. oralis*, *S. mitis*, and *S. sanguinis* (25). PI-2 pili are present in a limited number of *S. pneumoniae* strains (26, 27) and facilitate adhesion to eucaryotic cells. A PI-2 islet is present in *S. oralis* Uo5, and we used the deduced PitB protein, the major pilus subunit, to screen the primate genomes for the presence of pilus variants. Six primate genomes contained *pitB*-related genes, five *S. oralis* genomes were from a variety of primates, and one was of unclear species. All of them encoded PitB variants distinct from that of the reference strain *S. oralis* ATCC 10577 used by Zähner et al. (25) (Fig. 5). All these data indicate that interspecies gene transfer is a common feature among viridans streptococci independently of the source of isolation; alternatively, the genes have been lost in some of the strains.

In further analyses, we concentrated on cell surface components, namely, genes encoding enzymes for peptidoglycan and teichoic acid biosynthesis, cell surface proteins, and virulence factors not detected in *S. mitis* B6 (3) or *S. oralis* Uo5 (28), since they are the major factors responsible for the interaction with host cells.

Penicillin-binding proteins and MurMN. *S. pneumoniae* contains six PBPs (the bimodular transglycosylase/transpeptidases PBP1a, -1b, and -2a, the transpeptidases

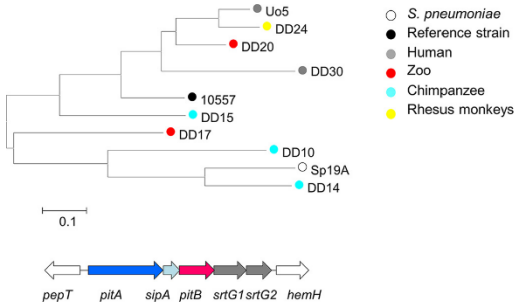


FIG 5 Comparison of PitB genes deduced from pilus clusters in streptococcal genomes. The phylogenetic tree was generated with MEGA6 using the Clustal alignment. PitB of *S. pneumoniae* SpTCH8431/19A and *S. oralis* ATCC 10557 represent the closest tBLASTn matches to PitB from *S. oralis* Uo5. The pilus cluster of *S. oralis* Uo5 is shown below. Red, *pitB*. The bar indicates changes per nucleotide position. *pitA* and *pitB*, pilus proteins; *sipA*, essential for pilin biosynthesis; *srtG1* and *srtG2*, sortases. White genes mark conserved flanking genes outside the pilus cluster.

PBP2x and -2b, and the D,D-carboxypeptidase [CPase] PBP3), and homologs to all six PBPs were present in the streptococcal genomes of this study. Resistance to β-lactam antibiotics is due to alterations in at least three PBP genes, PBP2x, PBP2b, and PBP1a, which are known to be encoded by mosaic genes in resistant *S. pneumoniae*, *S. mitis*, and *S. oralis* strains (for a review, see reference 29). No mosaic structures have been detected so far in PBP2a, which is only occasionally involved in strains of high-level resistance, and PBP3 is not known to contribute to resistance in clinical isolates of *S. pneumoniae*.

Genes encoding all six PBPs were found in all genomes analyzed here. Surprisingly, more than one CPase homolog was found in several of the streptococcal genomes (*S. constellatus*, *S. parasanguinis*, *S. gordonii*, *S. sinensis*, and *S. cristatus*), which formed two well-separated homology clusters (group 1 and group 2 in Fig. 6). The larger group, group 1, which roughly reflects the phylogeny of the species, consisted of the common PBP3 homolog present in all genomes, as expected for a gene product of the core genome, whereas this is less obvious for group 2 CPases. All group 2 CPases contained the active-site motifs SMSK, SSN, and KTG; *Streptococcus* sp. strain DD10 even contained a second group 2 protein with the deduced motifs SMAK, SSA, and KTG. This indicates that all group 2 proteins are functional enzymes. In contrast, in almost all strains with group 2 CPases, with the exception of *S. constellatus* strain DD09, the group 1 CPases had mutations at the active-site serine and/or at the conserved lysine residue within the SXXK motif (strains with this motif are marked with an asterisk in Fig. 6), suggesting an inactive enzyme. In all genomes where a group 2 CPase was present, the PBP3 homolog was positioned between the SufF gene and a gene encoding an ABC transporter. In contrast, group 2 CPases were found at three different genomic environments, indicating that they were acquired later during evolution.

The three PBPs known to be related to the resistance phenotype (PBP2x, PBP2b, and PBP1a) were examined more closely. The aims were to see how variable PBP sequences are among the *S. oralis* isolates, whether PBP sequences are shared between PBP genes from human and primate isolates, and whether signs of gene transfer are detectable.

Mosaic structures were apparent in penicillin-sensitive isolates from primates compared to the penicillin-sensitive strain *S. oralis* ATCC 35037 (Fig. 7), in agreement with the high variability of PBP genes detected in a large number of human commensal streptococci (30). Interestingly, the sequences from the three genes obtained from Taï chimpanzees were distinct from those of all other PBP genes; BLAST searches of sequences in the NCBI data bank also did not reveal any identical genes. This confirms that these isolates belong to a special group of *S. oralis* strains that has evolved

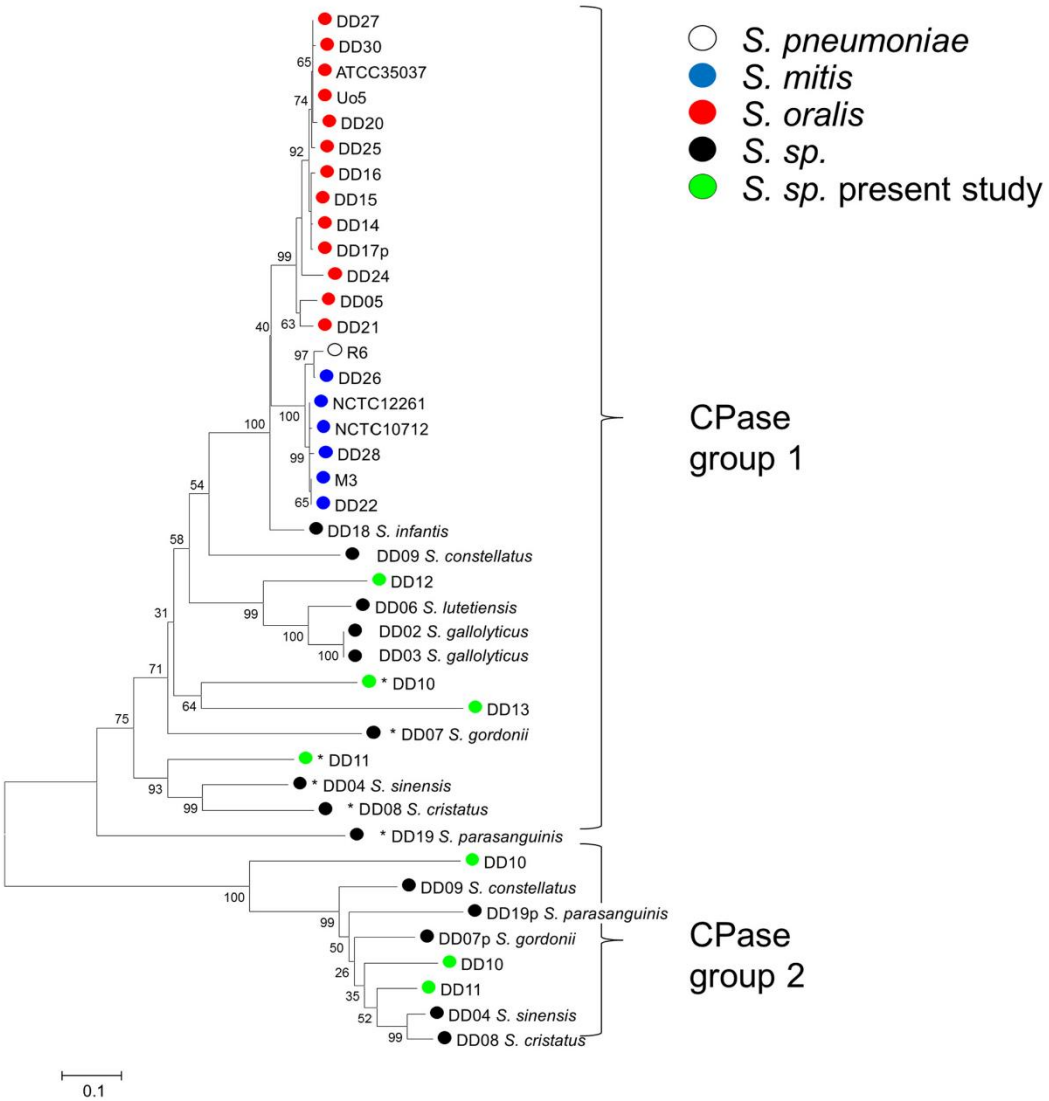


FIG 6 Distribution of PBP3 homologs in streptococcal genomes. A phylogenetic tree was constructed from deduced protein sequences from PBP3 D,D-carboxypeptidases using the MEGA6 software and muscle alignment. Proteins with at least one unusual active-site motif, indicating a nonfunctional PBP, are marked by asterisks, and partial sequences indicated by *P* bootstrap values (percentages) are based on 1,000 replications. The bar refers to genetic divergence as calculated by the MEGA software.

independently. Moreover, the primate isolates clustered separately from the human isolates, with the exception of *S. oralis* strain DD20 (bonobo isolate), which was closely related to ATCC 35037, and *S. mitis* strain DD22 (gorilla isolate), which carries PBP genes almost identical to *S. mitis* M3 genes (Fig. 7 and Fig. S2). These data also clearly indicate that there is no correlation between PBP2x, PBP2b, and PBP1a sequences; PBP2x from *S. oralis* strains DD14, DD15, and DD17 were identical, and all PBP2b and PBP1a sequences differed from each other (Fig. S2).

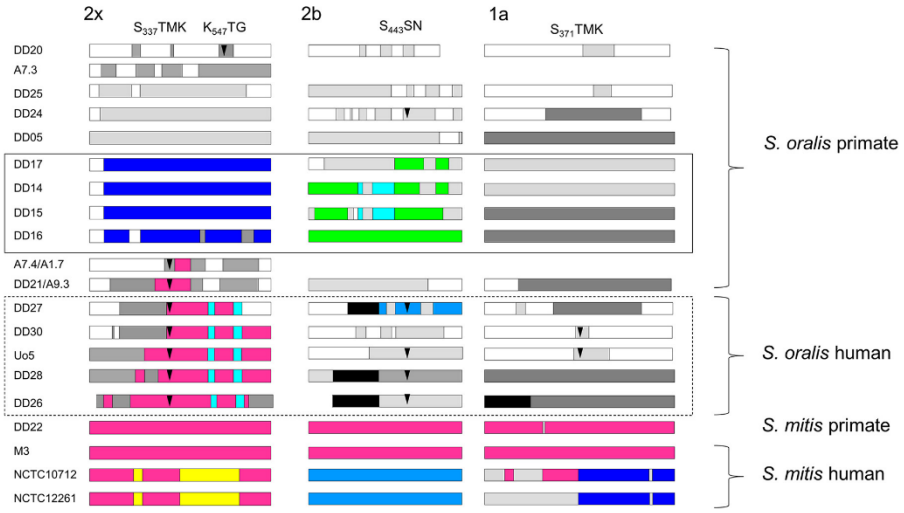


FIG 7 Mosaic PBPs in *S. mitis* and *S. oralis*. Mosaic gene structures were deduced by comparison to *S. oralis* ATCC 35037 PBP genes. Sequences that are highly similar to each other (<5% difference) are shown in the same color; sequences of different colors diverge from each other by at least 15%. Light gray, divergence from *S. oralis* ATCC 35037 of approximately 5%; dark gray, divergence by >15%; arrowheads, mutations within or close to active-site motifs which are shown on top; solid-line box, PBPs from free-living chimpanzees; dashed box, PBPs from human isolates with high-level penicillin resistance. The amino acid numbers are those of PBPs from sensitive *S. pneumoniae* strains.

Mosaic structures were most obvious in PBP2x, PBP2b, and PBP1a genes from the human isolates with high-level penicillin resistance (Fig. 7). The mosaic PBP2x genes belonged to the major PBP2x family common among oral streptococci, with a large sequence block highly related to some *S. mitis* strains (10). The mosaic structures indicate multiple gene transfer events among different species.

Mutations known to contribute to the resistance phenotype (for a review, see reference 29) were detected only within or close to the active-site motifs. Seven PBP2x variants contained T338A, whereas PBP2x of *S. oralis* DD20 contained the mutation Q552E, consistent with the lower susceptibility to cefotaxime of the strain. PBP2b mutations were also common (T446A), and *S. oralis* DD30 contained the same PBP1a mutation as *S. oralis* Uo5 (T372S) (see the arrowheads in Fig. 7). The PBP2b mutation confers only a small increase in β-lactam MICs (31), and therefore it is not surprising that *S. oralis* DD24 containing this mutation expresses only marginal resistance. In all cases, the PBP mutations were located within mosaic blocks; i.e., they have been acquired by gene transfer and are thus most likely not spontaneous mutations. In summary, mosaic structures are common also in *S. oralis* PBP2x genes, not only in the resistant isolates but also among penicillin-sensitive strains.

PBPs use muropeptides as the substrates for their transpeptidation reaction and the formation of cross-links in the peptidoglycan. In *S. pneumoniae*, MurM and MurN enzymes, which are responsible for the synthesis of branched muropeptides, have been described. MurM adds an L-Ala or L-Ser to the ε-amino group of the L-Lys residue of lipid II, and MurN adds another L-Ala residue. The branched peptides are used as an acceptor substrate for the transpeptidation reaction of PBPs, resulting in interpeptide bridges in mature peptidoglycan (32). MurM genes have a mosaic structure in some penicillin-resistant *S. pneumoniae* strains (33) and are thus also the subject of gene transfer events. We recently showed that *S. oralis* Uo5 contains an unusual MurM gene and no *murN*, consistent with the presence of branched muropeptides containing only 1 alanine residue attached to lysine (34) instead of the Ala-Ala or Ser-Ala dipeptide found

in *S. pneumoniae* (35–38). We therefore searched the genomes for the presence of *murMN* to see whether the lack of *murN* is a common feature of *S. oralis*.

Although MurMN was present in most streptococcal genomes (Table S2), the situation among the *S. oralis*/*S. mitis* group was surprisingly varied. We found MurM homologs only in two *S. oralis* genomes (DD17 and DD21) and not in the other *S. oralis* genomes, regardless of the MurM variant used in BLAST searches (MurM from *S. oralis* Uo5, *S. mitis* B6 or *S. infantis* DD19). Similarly, BLAST searches of *S. oralis* draft genomes in the NCBI data bank revealed only one genome which contained a MurM homolog. Also, the *S. mitis* genome DD22 contained *murM* within a genomic environment similar to that of *S. oralis* Uo5 but not *murN*, whereas other *S. mitis* genomes contained *murMN* in a genetic environment similar to that of *S. mitis* B6.

TAs and choline-binding proteins. The genomes were screened for genes required for teichoic acid (TA) backbone biosynthesis and decoration. The sequenced strains can be divided into two groups. The first group of 13 strains contains one gene whose product is the key enzyme LtaS, the lipoteichoic acid (LTA) synthetase which catalyzes the polymerization of type I LTA, containing a polyglycerolphosphate chain, the most frequently encountered cell wall polymer (39). This group includes different *Streptococcus* species (Table S2), and their entire LtaS proteins are similar to LtaS of *S. mitis* B6 (Smi0753), with 58% to 75% of their amino acids being identical to those of LtaS of B6. Three *S. mitis* strains (DD22, DD26, and DD28) contained a deduced LtaS protein with >96% identity to LtaS of B6. In contrast, *ltaS* homologs were not found in *S. oralis* and *S. infantis* genomes, in agreement with published data (40).

The 12 strains where *ltaS* was absent contained the genes involved in the biosynthesis of the unusually complex, choline-containing type IV LTA, typical for *S. pneumoniae* and closely related species. In *S. pneumoniae*, LTA and wall teichoic acid exhibit identical structures within their repeating units (RU) (41). In the *S. mitis* B6 strain, the TA gene content and genetic organization are nearly identical to those of *S. pneumoniae* R6, except that *S. mitis* may contain galactose instead of glucose in its TA repeating unit (40). In contrast, *S. oralis* Uo5 produces a structurally different TA repeating unit and has structural complexity even greater than that of pneumococcal LTA (42). Two *S. mitis* isolates, DD22 and DD28, contain *S. pneumoniae*-type TA biosynthesis clusters but differ in their glycosyl transferase genes, suggesting that DD22 contains glucose but that DD28 contains galactose in its TA. In contrast, all *S. oralis* isolates and *S. mitis* DD26 contain glycosyl transferase genes of the *S. oralis* Uo5 type; the *S. infantis* DD18 *licD4* cluster also differed from that of *S. oralis* Uo5. All three species contain the genes for uptake and activation of exogenous choline (*licABC*) as well as for decoration of teichoic acids (*licD* homologs). A closer look revealed that one group (*S. mitis* DD26 and *S. oralis* DD16, DD17, DD20, and DD21) contained a *lic4* region where the homology of the *licD3* and *tacF* gene products to the *S. oralis* Uo5 proteins was much lower (Table S2). This suggests that at least four biochemical variants of choline-containing teichoic acids occur in *S. mitis*, *S. oralis*, and *S. infantis*. The results are in agreement with data obtained by Kilian et al. showing that monoclonal antibodies directed against the backbone and the phosphocholine residues of TAs react with some strains of these three species (6).

CBPs. Choline-binding proteins (CBPs) are anchored to the cell wall by hydrophobic interactions with choline-containing teichoic acids (for a review, see reference 43). They are composed of a choline-binding module consisting of repeats of 20 amino acids and a nonconserved functional domain. They represent a highly varied family with respect to non-CBP modules, and numbers of CBPs also vary largely even within one species.

There are only three CBPs common to *S. mitis* B6, *S. oralis* Uo5, and *S. pneumoniae*, namely, LytB, a key enzyme for cell separation (44), CbpD, a murein hydrolase implicated in the lysis of noncompetent genes (45), and CbpF, a putative modulator of cell wall hydrolases (46), strongly suggesting that these CBPs have an important physiological role in these species. Genes encoding these three CBPs were found in the *S. oralis*, *S. mitis*, and the *S. infantis* genomes, which contained the *lic* clusters described above. All streptococcal genomes that did not contain CBPs, and thus did not contain

CbpD, encoded a protein related to LytF of *S. gordonii* and possessing a similar function (47, 48) or another, new autolysin with a related CHAP domain (*Streptococcus* sp. strains DD10 and DD13).

Noncoding csRNAs 1 to 6. Streptococci contain a two-component regulatory system, CiaRH, which was originally identified in and characterized for *S. pneumoniae* (49–51). The system affects diverse processes, such as genetic competence (52, 53), bacteriocin production (54), host colonization (55), virulence (56), and β-lactam resistance (49, 57). Within the CiaR regulon, there are variable numbers of genes specifying small noncoding RNAs, i.e., *cia*-dependent small RNAs (csRNAs), ranging in size from 51 to 200 nucleotides (59). Since the csRNAs are involved in major CiaR-associated phenotypes in *S. pneumoniae* (53, 58), it was of interest to look for csRNA genes in the novel genomes.

First, the genomes were searched for the response regulator CiaR, which was clearly detected in all genomes with the typical recognition helix described previously (59). The corresponding histidine kinase, CiaH, was also present and showed a greater variability than CiaR, consistent with an earlier observation (59). Subsequently, the genomes of *S. mitis*, *S. oralis*, *S. gallolyticus*, and *S. gordonii* strains were searched for the types of csRNAs previously defined in other strains of these species (59). All csRNAs were present in the new strains. Interestingly, some *S. oralis* strains contained six instead of the five csRNAs of *S. oralis* Uo5 (17) caused by duplications of csRNA2, csRNA4, or csRNA6 genes. Two species with unknown repertoires of csRNAs contained csRNAs known from other species. *S. lutetiensis* harbored the *S. gallolyticus* UCN34 csRNAs except for csRNA40 (59), and *S. infantis* harbored four of the five *S. oralis* Uo5 csRNAs but not csRNA1. The other streptococci, especially those without species designation, did not yield full-length hits in the BLAST analysis with csRNA types defined by Marx et al. (59), indicating the existence of novel csRNAs in these bacteria.

Closer inspection of *S. oralis* sequences with duplicated csRNA genes revealed a surprising result. In between duplicated csRNA2 and csRNA6 genes (DD05 and DD15), we found a genetic island of four genes encoding redox proteins related to succinate dehydrogenase and fumarate reductase, a transporter of the oxalate/formate antiporter family and an AraC-type regulator. These genes are not present in *S. oralis* strains without duplicated csRNA genes. It appears therefore, that this small metabolic island is integrated into the *S. oralis* genome via csRNA genes. Similarly, an even smaller island of two genes encoding proteins without assigned functions is integrated between duplicated csRNA4 genes (DD27).

In *S. infantis* DD18, the four-gene island of *S. oralis* DD05/DD15 is integrated between two csRNA6 genes. A phage is apparently integrated into csRNA2 in DD14, but we could not deduce whether this is also related to a csRNA gene duplication due to termination of the contig sequence.

***S. pneumoniae* virulence factors in viridans streptococci.** A large number of surface components important for the interaction with host cells have been described to occur in *S. pneumoniae* (for reviews, see references 60 and 61). Most of these genes were present in all genomes of this study, as has been described for *S. mitis* B6 (3), including the lipoprotein PsaA, a manganese transporter, and the two peptidyl-prolyl isomerases SlrA and PpmA, as were the nonclassical cell surface proteins, the plasminogen-binding proteins GAPDH (glyceraldehyde-3-phosphate dehydrogenase) and enolase, and the fibronectin-binding protein PavA. PavA is essential for colonization in the upper respiratory tract but probably mediates adherence indirectly by affecting other virulence factors (62, 63). The high conservation of PavA is exemplified in Fig. S3. Despite a high degree of sequence identity, every genome contained a distinct predicted PavA which differed from *S. pneumoniae* PavA by up to 3.6% (*S. mitis*) and 5.4% (*S. oralis*). In this context, it is interesting that none of the viridans streptococci investigated here contained the gene cluster implicated in riboflavin biosynthesis (*S. pneumoniae* R6 spr0161 to spr1064) except the *S. gallolyticus* and *S. lutetiensis*

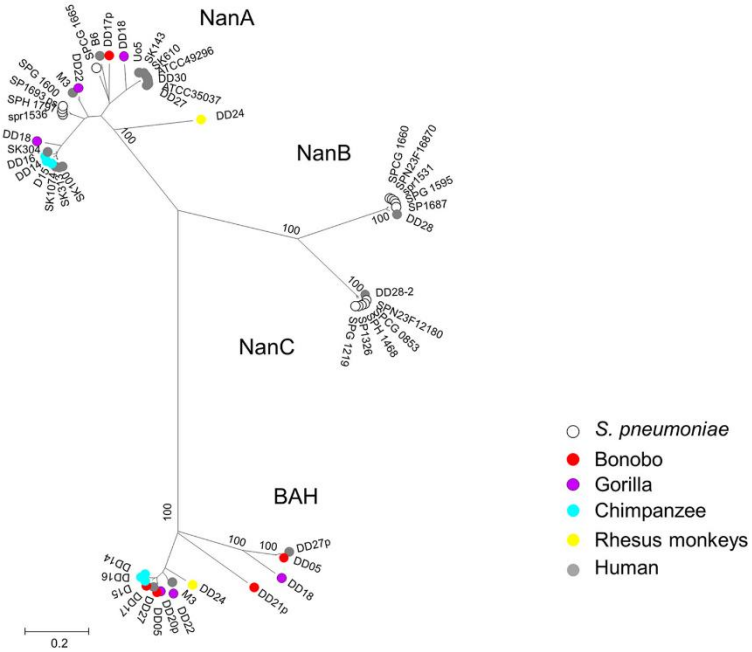


FIG 8 Neuraminidases and a conserved *β*-*N*-acetyl-hexosaminidase in *S. pneumoniae*, *S. mitis*, and *S. oralis*. The phylogenetic tree was generated with MEGA6 using the muscle alignment from the genomes used in this study and selected *S. pneumoniae* genomes. NanA, NanB, and NanC are indicated. p, partial sequences; BAH, putative *β*-*N*-acetyl-hexosaminidase. Bootstrap values (percentages) are based on 1,000 replications. The bar refers to genetic divergence as calculated by the MEGA software.

genomes. In contrast, the thiamine cluster absent in *S. mitis* B6 (3) was variably present in several *S. oralis* genomes (see Table S2 in the supplemental material).

We investigated neuraminidases in more detail, since these enzymes target sialic acids, which differ between humans and primates. *N*-Acetylneuraminic acid (Neu5Ac) and its derivative *N*-glycolylneuraminic acid (Neu5Gc) are major sialic acids in many vertebrates, including the great apes. However, Neu5Gc is missing in human tissues due to an inactive form of the enzyme required for the generation of this compound (64, 65). In *S. pneumoniae*, three neuraminidases have been described: NanA, which contains an LPXTG motif, and NanB/NanC (for a review, see reference 4). NanA contributes to attachment to host cells by hydrolyzing terminal sialic acid residues from host proteins and polysaccharide components. The *S. mitis* genome of DD22 and many of the *S. oralis* genomes contained a closely related *nanA* homolog, as did the genomes of DD08 (*S. cristatus*), DD10 (unknown species), and DD04 (*S. sinensis*) (Table S2). In contrast, NanBC were absent in all *S. oralis* isolates and found only in one human *S. mitis* isolate, DD28 (Table S2). Instead, a protein encoding a *β*-*N*-acetyl-hexosaminidase occurred in most oral streptococci (Table S2), which was absent in all *S. pneumoniae* genomes. Like NanA, it contains a YSIRK signal peptide and represents an LPXTG cell surface protein. No primate-specific clustering was observed (Fig. 8).

A few genes described as *S. pneumoniae* virulence factors were not detected in the *S. mitis* B6 (3) or the *S. oralis* Uo5 (28) genome. This includes the three CBPs PspA, PcpA, and PspC (for reviews, see references 60 and 66), the hyaluronidase HysA, and the *cps* cluster responsible for the highly variable polysaccharide capsule. As shown recently, *cps* clusters have been imported from numerous *Streptococcus* species (5) and were not investigated here. Hyaluronidase activity is present in most *S. pneumoniae* isolates. It

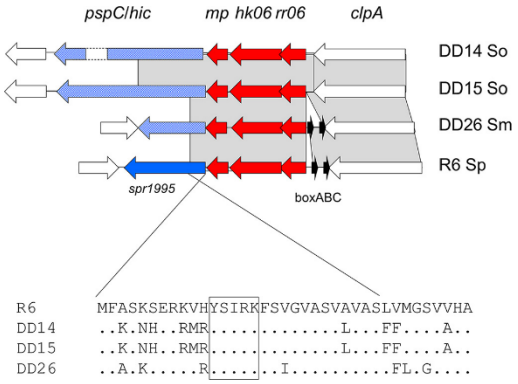


FIG 9 TCS06 islet. The genes encoding TCS06 and a conserved putative membrane protein (mp; red arrows) and a cell surface protein (blue arrow, PspC in *S. pneumoniae* [Sp] R6; blue hatched arrows, an LPXTG-containing protein in *S. mitis* [Sm] DD26 and *S. oralis* [So] DD14 and DD15) are shown. White genes represent flanking regions. Gray areas mark BLASTn matches between the sequences. BoxABC elements are shown in black. The signal peptide of the proteins is shown below the diagram; the pentapeptide motif conserved in many cell surface proteins is boxed.

has been found in some *S. oralis*, but not in *S. mitis*, strains (6). Consistently with this observation, only one *S. oralis* genome harbored a HysA gene (DD25) (Table S2).

PspA is a highly immunogenic protein, and antibodies against PspA protected mice when challenged with *S. pneumoniae* (60, 66). It interferes with complement activation and is able to bind lactoferrin (67, 68). PspA sequences are highly divergent in *S. pneumoniae* due to intragenic recombination similar to that of PspC (60, 69). PspA has a mosaic structure in its central highly charged and proline-rich regions. BLAST analysis performed with the non-choline-binding domains revealed that only in the human isolate *S. mitis* DD28 is a PspA-related deduced protein identical to *S. mitis* B6 CBP2 (smi0038) present. N- and C-terminal sequences were closely related to *S. pneumoniae* PspA; however, the charged and proline-rich regions were distinct.

PcpA is conserved among pneumococci, and since it elicits protection in murine models of pneumonia and sepsis, it is now included in vaccination trials (70). The PcpA gene is associated with transposase elements, indicating acquisition from a still-unknown source. We found only in DD09 (*S. constellatus*) a PcpA homolog with 72% identity to *S. pneumoniae* PcpA (first 360 amino acids). However, it lacked the choline-binding domain and carried an LPXTG motif instead. We did not find evidence in other *S. constellatus* genomes for the presence of this gene, strongly suggesting that it is part of the accessory genome in this particular case.

S. pneumoniae PspC (also named CbpA) interacts with the secretory component of the polymeric immunoglobulin receptor and interacts with components of the innate immune system, such as the complement proteins C3 and factor H (71); it also binds to vitronectin (72). It is located on an island encoding TCS06 and an integral membrane protein of unknown function. Some isolates contain another PspC-like protein, but one which shows a C-terminal cell wall-anchoring LPXTG motif instead of the choline-binding repeat, and differ also in their proline-rich domains. PspC was also named Hic for factor H-binding inhibitor of complement (73, 74). PspC/Hic proteins are highly varied in *S. pneumoniae*, and only the signal peptide, as well as the overall domain organization, is conserved (74).

PspC homologs containing the highly conserved N-terminal signal peptide of *S. pneumoniae* PspC were found in *S. oralis* DD14 and DD15 as well as in *S. mitis* DD26, encoded by a gene located in the same genetic environment as in *S. pneumoniae*, downstream of a TCS06 homolog (Fig. 9). However, as with Hic, these proteins con-

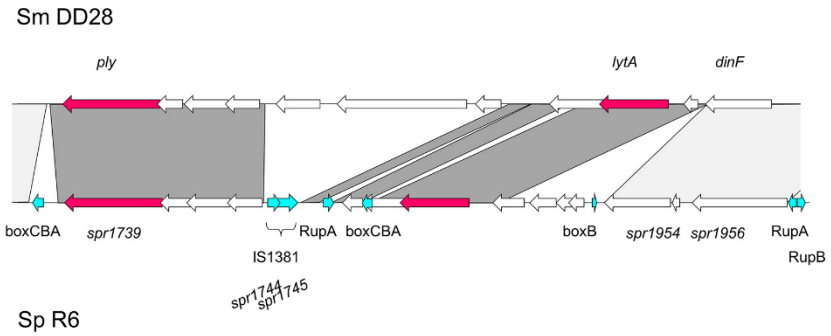


FIG 10 *ply-lytA* **island in** *S. mitis* **DD28. Shown is a comparison of the region containing** *ply* **and** *lytA* **between** *S. mitis* **DD28 and** *S. pneumoniae* **R6. The genes** *ply* **and** *lytA* **are shown as red arrows. Dark-gray areas mark BLASTn matches between the two regions; light gray marks regions conserved in other strains of these species. Turquoise arrows, Rup, BOX, and IS elements.**

tained an LPXTG motif and no choline-binding module and differed largely in their proline-rich internal regions; both *S. oralis* proteins were almost identical throughout the first 319 residues. It should be noted that the *pspC* island in *S. pneumoniae* is frequently associated with BOX elements (strain R6) or transposases (strain Hungary 19A_6) which may be involved in the variability of this region. Such elements were missing in the *S. oralis* pspC islands, and only two BoxABC elements were present in the *S. mitis* DD26 genome.

LytA autolysin and Ply pneumolysin. In contrast to all other CBPs, LytA does not contain a signal peptide and is therefore located mainly in the cytoplasm of the cells (75–77). It is still unclear how it accesses the pneumococcal cell wall, and it has been suggested that its activity is restricted to sites of nascent peptidoglycan biosynthesis (78). LytA encodes the major autolysin in *S. pneumoniae* and is responsible for stationary-phase lysis of pneumococcal cultures and for the lytic response to β-lactams and other cell wall inhibitors. It acts during genetic competence to lyse noncompetent cells, a process named "fractricide" (79), and it is probably required for the release of virulence factors, including the pneumolysin Ply (80). In *S. pneumoniae*, the LytA gene is located on a genomic islet, including the Ply gene, and has been imported probably via recombination with phages, which frequently carry a *lytA* homolog (81). LytA genes associated with cryptic phage relicts are genetically distinct. The presence of *lytA* and *ply* in *S. mitis* has been documented (3, 6, 81–83), but their genomic organization has not been elucidated.

LytA homologs were found in the genomes of three *S. oralis* isolates from primates. LytA from a wild chimpanzee was closely related to LytA from a zoo ape. Two human *S. mitis* isolates contained two copies, one of which was associated with phage genes, whereas the other one was located downstream of *dinF*, as with *S. pneumoniae* lytA. Both *S. mitis* DD28 and DD26 also contained a *ply* homolog. The organization of *lytA-ply* in *S. mitis* DD28 was similar to that in *S. pneumoniae* but included two genes not present in *S. pneumoniae*, while *S. pneumoniae* contained multiple fragmented insertion sequence (IS) elements and other mobile sequences, such as RUP and BOX elements (Fig. 10). The rare occurrence of RUP elements among *S. mitis* strains has been noted (5). A similar organization can be deduced from the DD26 genome, but it contained a sequence gap between *lytA* and *ply* (not shown). This shows that the complete island is present also among the *S. mitis* strains from humans.

In summary, almost all genes associated with *S. pneumoniae* virulence were found in the primate isolates. However, a small set of genes encoding the PspC islet, HysA, NanBC, and Ply-LytA, present in most *S. pneumoniae* strains, were restricted to a few *S. oralis* and/or *S. mitis* genomes.

DISCUSSION

Species isolated from primates. One part of this study was to see which primate species contained streptococci related to those that are commensals in human. Streptococci could be isolated from great apes, Old World monkeys (rhesus monkeys), and lemurs (ring-tailed lemurs and Verreaux's sifaka) and included a wide variety of species of viridans streptococci according to MLSA and genomic analyses. No streptococci of the Mitis group of viridans streptococci could be isolated from lemurs from Madagascar; we found only *S. gallolyticus*, which belongs to the Bovis group. However, members of the Mitis group of streptococci were obtained only from monkeys and great apes, with *S. oralis* being the predominant species. None of our samples showed 16S rRNA identity to other species of the Mutans group, including *Streptococcus troglodytes*, *Streptococcus dentirousetti*, *Streptococcus downei*, and *Streptococcus macacae*, which have been isolated from plaques of chimpanzees obtained by brushing their teeth (84), probably due to the different methods used for sampling. The streptococci included novel species isolated from wild chimpanzees that we could not define by MLSA or 16S rRNA analysis, and other genes did not reveal matches with >90% homology (Table S3). These findings should be corroborated by microbiome analyses to reveal potential differences between the commensal floras of primates.

The genomes of three strains isolated from ring-tailed lemurs defined one strain, KG3c, closely related to *S. dysgalactiae* subsp. *equisimilis* and two strains of the Mitis group, *S. lutetiensis* DD06 and *S. sinensis* DD04. It was difficult to obtain MLSA sequences from other ring-tailed lemur isolates, and only sequences from *rpoA* and *pyk* could be obtained from another seven strains. Phylogenetic analysis showed that they formed an unidentified cluster between *S. sinensis* and *S. gordonii* (not shown). *S. oralis* strains were obtained only from Old World monkeys, most of which were located on branches distinct from those containing the main cluster of human isolates in the phylogenetic trees generated by MLST and MLSA. This strongly suggests that *S. oralis* had evolved in these animals prior to the origin of humans and that this species is part of the commensal flora at least of great apes. The finding that *S. oralis* is also associated with rhesus monkeys should be confirmed by screening of wild animals, since the possibility of transfer of strains from humans to animals held in captivity cannot be excluded. Taken together, the phylogenetic tree of viridans streptococci appears to parallel the evolution of primates. Obviously, more samples from free-living animals, including New World monkeys, are needed to gain a comprehensive view of the evolutionary history of streptococci, which are important commensals in these animals.

In this context, it is remarkable that antibiotic resistance phenotypes and the TetM resistance determinant were found only in isolates from zoo animals and in those from the German primate center, i.e., in an environment where these phenotypes are frequent among *Streptococcus* spp. but not in free-living animals. Nevertheless, the possibility of transfer from humans to wild animals cannot be excluded. The sampling of fruit wedges is a successful strategy to screen for bacteria and viruses (85), and a human-to-monkey transmission of *S. aureus* has been reported (86).

As pointed out before (5, 8), the highly diverse subclusters within the *S. mitis* cluster could not be distinguished by phenotypic properties, challenging the definition of species. This is also apparent if one considers the *S. oralis* lineages (Fig. 2), which are probably a reflection of the diversification in different nonhuman hosts. Remarkably, the newly defined species *S. tigurinus* (87) and *S. dentisani* (15) cluster among the organisms of the *S. oralis* subcluster of IgA protease-negative organisms and the previously defined *S. mitis* biovar 2 subcluster (Fig. 2A), respectively, challenging the definition of *S. oralis*. 16S rRNA sequences are varied within designated *S. oralis* isolates, and thus the species of isolates that cluster within the heterogeneous *S. oralis* cluster, including the genomic information of many more strains, should be confirmed by further analyses. Determination to the species level is aggravated by the capacity for genetic transformation in viridans streptococci. The large accessory genome bears many signs of interspecies gene transfer, including large genomic islands common

among different streptococcal species, leading to a smooth transition between species in comparative genomic hybridization experiments of oral streptococci (3, 4, 88). In this context, it is remarkable that csRNA genes apparently serve as entry sites for horizontal gene transfer in several cases, as described here, thereby contributing to the genomic variability observed for *S. oralis* and resulting in an overlap in the accessory genomes of *S. oralis* and *S. infantis*. It will be interesting to see if csRNAs with inserts are also found or are found even more often in other streptococcal species.

Genome analysis of peptidoglycan and teichoic acid biosynthesis. The second part of this study investigated cell surface components, including enzymes involved in cell wall polysaccharide biosynthesis, extending previous genomic analyses that focused mainly on *S. mitis* (3, 5). The variability of genes involved in peptidoglycan biosynthesis among *S. oralis* strains—those for PBPs and MurMN—is astounding. A high variability of PBPs in *S. mitis* is well known and has been exemplified recently using a large number of isolates (30). We now provide evidence that *S. oralis* isolates also differ largely in sequences encoding PBP2x, PBP2b, and PBP1a, proteins implicated in β-lactam resistance, and that these proteins are known to have a mosaic structure in resistant isolates. Mosaic blocks present in a common class of resistant mosaic PBP2x genes that are closely related to PBP2x genes from sensitive *S. mitis* isolates were found only in human isolates with high-level resistance (Fig. 7; see Fig. S2 in the supplemental material). In contrast, PBP2a sequences were conserved throughout the sequences (not shown). In several genomes, two genes encoding CPase PBP3 homologs, termed group 1 and group 2 CPases, were present (Fig. 6). In most cases, only the group 2 CPase gene encoded a protein with conserved active-site motifs. It is likely that that the products of these genes represent the only functional CPase. The gene encoding group 1 CPases was located in the same genomic environment, whereas the genes encoding group 2 CPases were present at locations in DD04 (*S. sinensis*), DD08 (*S. cristatus*), and DD07 (*S. gordonii*) that were distinct from those in DD09 (*S. constellatus*), DD11 (unknown species), and DD19 (*S. parasanguinis*); the genes in DD10 (unknown species) were again positioned differently, suggesting that these genes have been imported into the genomes on different occasions. It would be interesting to see whether the enzymatic activities of group 2 enzymes differ from those of group 1 enzymes.

S. oralis Uo5 lacks MurN, associated with an interpeptide bridge consisting of only one L-Ala residue (34), and we now show that *murM* and *murN* are also apparently lacking in some isolates (see Table S2 in the supplemental material). Accordingly, these strains most likely contain no interpeptide bridges in their peptidoglycan, which should be confirmed by biochemical analyses. This raises the question of which of the PBPs is preferentially affected by an altered substrate, a question that can be clarified only by complex genetic or biochemical experiments. Jensen et al. also noted the absence of MurM homologs in *S. mitis* and *S. oralis* (30) and hypothesized that this genotype is tolerated only in penicillin-sensitive strains. Interestingly, we found mutations in PBPs associated with resistance (Q552E in PBP2x of DD20 and T446A in PBP2b of DD24) in *S. oralis* isolates where both *murM* and *murN* were missing. Since deletion of MurMN in penicillin-resistant strains leads to a breakdown of resistance, including cefotaxime resistance, which is mediated by PBP2x but not by PBP2b, it has been speculated that it is the altered "resistant" PBP2x whose function depends on the presence of branched peptides (89). Given the variability of PBP sequences and of PBP2x in particular, it is quite possible that resistant PBP variants that are still functional even in the absence of MurMN have evolved. However, it might be difficult to find such isolates. Since resistant PBPs evolved in the genomic context of the respective *murMN* constellation and are transmitted mainly by gene transfer, resistant *de novo* variants of strains that do not contain *murMN* might be encountered only on rare occasions.

The variation observed in *S. oralis* strains with respect to teichoic acid biosynthesis clusters responsible for choline decoration of teichoic acids (LTA type IV) is astounding. All strains contained genes required for choline-containing TAs, but the presence of distinct *lic* clusters (*lic3* versus *lic4* of Uo5 and *lic4* of *S. oralis* cluster 2 isolates) strongly

suggests three different biochemical makeups of this cell surface polysaccharide in *S. oralis* strains and in at least two variants of *S. mitis*. Kilian et al., using monoclonal antibodies to detect epitopes characteristic of the backbone and the phosphocholine residues of the TA, showed that *S. infantis* contains choline in its cell wall (6). We now provide genetic evidence for the presence of these components in *S. infantis*, with DD18 containing a *licD4* cluster similar to the *licD4* cluster in Uo5; more *S. infantis* genomes are needed to confirm that *lic* genes and CBP genes are uniformly present in this species and whether variants occur, as shown here for *S. oralis*. All species with type IV LTAs share the physiologically important choline-binding proteins CbpD, CbpF, and LytB. The other viridans streptococci investigated here contain LtaS synthase to polymerize a much simpler LTA (type I) consisting of a polyglycerolphosphate chain (39). Some *S. mitis* strains contain *ltaS* in addition to the *lic* operons, similar to what occurs in *S. mitis* B6 (40). As pointed out before, experimental evidence is required to know whether these strains express two types of LTA.

***S. pneumoniae* virulence factors.** In general, this study confirmed that many genes encoding so-called virulence factors of *S. pneumoniae* are present in many strains among viridans streptococci, as has been shown in several genomic studies (3, 5, 6, 28). We focused our analysis on cell surface proteins since these are the components that interact with host cells and thus are potential candidates to reveal differences between *S. pneumoniae* and related streptococci. The main finding was that the neuraminidases NanBC, which are variably present in *S. pneumoniae* genomes, were found only in one *S. mitis* isolate and were completely absent in *S. oralis* and other viridans streptococci. In contrast, a related protein with predicted *N*-acetyl-hexosaminidase activity occurred in all *S. oralis* genomes, independently of the host of the isolates, and was present also in the *S. infantis* genome. The *in vivo* role of this protein remains to be clarified. We failed to detect features that are exclusively associated with primate versus human isolates for several reasons. First of all, the sample size for one streptococcal species from a single primate species is still too small to interpret results reliably in this respect. It is possible that the capsule plays an important role for host specificity, as has been pointed out for *S. pneumoniae* (5), but due to the variability of capsular clusters, it is difficult to interpret the variability encountered, e.g., in *S. oralis* genomes. Also, differences that are due to host specificity might not be visible at the genomic level but require physiological tests or biochemical analyses (e.g., tests for the glycosylation pattern of surface components).

Two clusters which included *S. pneumoniae*-specific virulence genes were found among *S. mitis* and *S. oralis*: the *ply* (or *lytA*) gene cluster and the TCS06 *pspC* islet. The presence of *ply-lytA* in commensal streptococci is well known (3, 6, 81–83), but this is the first time that we can show that the entire island is present in some *S. mitis* strains and that it is located at the same genomic position as in *S. pneumoniae*. The main difference is the absence of the repeat elements RUP and BOX (Fig. 10). RUP elements have apparently undergone extensive expansion during the evolution of *S. pneumoniae* (90), whereas they are rarely found in *S. mitis* or *S. oralis* (3, 5). Similarly, the TCS06 *pspC* cluster (Fig. 9) includes BoxABC elements in *S. pneumoniae* which were missing in the two *S. oralis* genomes containing this islet; they were present in the *S. mitis* genome. Generally, BOX elements are much rarer in *S. oralis* than in *S. mitis* or *S. pneumoniae*. Sequences related to the three novel *S. mitis*/*S. oralis* PspC-like proteins were found in *S. pneumoniae* genomes (e.g., strains NT_110_58 and Hungary 19A_6), documenting a remarkable example of domain shuffling and protein diversification during evolution.

There are several open questions that remain. What is it that makes *S. pneumoniae* a pathogen? Do the *S. oralis*/*S. mitis* strains that contain PspC, HysA, and the *lytA-ply* island have a higher-pathogenicity potential than those that lack these components? Is it the combination of these well-known virulence genes plus PcpA, PspA, and the highly variable polysaccharide capsule (which are present in most pneumococcal strains) what imparts pathogenicity? What is the role of the *N*-acetyl-hexosaminidase in *S. oralis* and *S. infantis*? What is the driving force behind the variation observed in peptidoglycan

and the teichoic acid biosynthesis enzymes, PBPs, MurMN, and LicD3/4? Are there host-specific components that occur in human as well as in primate isolates? The speed of genomic research and novel biochemical tools might help to solve some of these riddles.

MATERIALS AND METHODS

Bacterial strains. Swabs were obtained from the Frankfurt Zoo (throat swabs from bonobos, orangs, and gorillas) and from the German Primate Center, Göttingen, Germany (throat swabs from rhesus monkeys and nose swabs from ring-tailed lemurs). Throat swabs from free-living lemurs (Verreaux's sifakas, *Propithecus verreauxii*; red-fronted lemur, *Eulemur rufifrons*; Western fat-tailed dwarf lemur, *Cheirogaleus medius*; gray mouse lemur, *Microcebus murinus*) in the Kirindy Forest in Madagascar, which is part of a field site operated by the German Primate Center, were obtained during a survey of anesthetized animals in the course of an annual marking and survey mission that followed the protocol described previously (91). Samples from wild chimpanzees from the Taï National Park, Ivory Coast, where contact to humans is highly restricted, were isolated from fruit wedges containing the fruit of two species of plants (memecylon and Parinari), which are chewed by the animals for long time periods and sucked on intensively before they are spit out. These fruit wedges where collected once the chimpanzees had reached a minimum distance of 10 m from the sample, which was placed in STGG medium (92) and transported to the field camp, where they were preserved in liquid nitrogen and shipped to Germany as described previously (86). Samples were vortexed and streaked on blood agar plates using a 10-μl inoculation loop. Plates were incubated overnight, and colonies showing alpha-hemolysis were isolated and tested for optochin susceptibility (see Table S1 in the supplemental material). Bacterial samples from swabs were grown in C medium (93) supplemented with 0.1% yeast extract, diluted, and streaked on D-agar plates (94) with 3% defibrinated sheep blood. Individual colonies suspected of representing viridans streptococci were isolated, and antibiotic susceptibility was tested with the Etest (β-lactam antibiotics) and antibiotic discs (all other antibiotics) (Table S1).

Bacterial genomes. The 25 genomes of isolates from primates and their accession numbers are listed in Table S1 in the supplemental material. In addition, seven genomes from human *S. mitis* and *S. oralis* isolates which were used in previous studies (10) were included for comparison (Table S1). Files with sequence reads from 454 3K paired-end sequencing technology were available for 26 strains isolated from various monkeys and monkey groups, including 24 from *Streptococcus* spp. The gsAssembler (Newbler), version 2.6, from Roche was applied for assembly. The rapid annotation subsystem technology (RAST) server (95) designed for annotation of bacterial and archaeal genomes was applied to obtain EMBL-formatted files containing protein, tRNA, and rRNA annotations from a large set of several output formats; *S. mitis* NCTC10712 was annotated by best BLAST analysis (96).

DNA isolation and PCR amplification. Chromosomal DNAs from streptococci were isolated as described previously (97). PCR products were purified using a JetQuick DNA purification kit (GenoMed). PCRs were performed using either Goldstar Red *Taq* polymerase (Eurogentec) or DreamTaq polymerase (Fermentas) according to the manufacturer's instructions. The oligonucleotides used in this study were obtained from Eurofins. PBP2x gene fragments were amplified with the primers pn2xup and pn2xdown, as described previously (97). 16S rRNA sequences were amplified by PCR with the bacterium-specific primers rRNA2 (TCAGATTGAACGCTGGCGGC) and rRNA1 (TATTACCGCGGCTGCTGGCA) or the *Streptococcus* sp.-specific primer rRNA-Strep1rev (CTTACGGTTACCTCACCGACTTCG) and rRNA2.

Identification of csRNA genes. The genomes were searched for csRNA genes by BLAST analysis using the genes of 40 csRNA types defined by Marx et al. (59) as a query. Hits covering at least 50 consecutive base pairs were taken, and their genomic upstream regions were visually inspected for the presence of a typical CiaR-regulated promoter with the CiaR-binding site NTTAAG-5-content-type="gene">TTTAAG placed 10 bp upstream of a −10 region. In all cases, such a promoter sequence was identified. It allowed us to predict exactly the start of the csRNA genes. Subsequently, the last T in the terminator region was taken to define csRNA genes completely.

Bioinformatic tools and analysis. BLAST searches were performed using the NCBI microbial genome data bank. A specialized search for the primate genomes was established on the NBC11 bioinformatic computational site http://nbc11.biologie.uni-kl.de/ (database searches/BLAST primate isolates) for all contigs. Neighbor-joining trees were generated with MEGA6.06 (98) and Clustal alignments using standard parameters; in some cases, muscle alignment was chosen, as stated in the text. Bootstrap analysis was based on 1,000 replicates; for MLST-derived phylogenetic trees, 500 replicates were used. For comparison, analyses were also conducted with the neighbor-joining algorithm.

Nucleotide sequence accession numbers. The whole-genome shotgun project sequences have been deposited in DDBJ/EMBL/GenBank (accession numbers are listed in Table S1 in the supplemental material). The versions described in this paper are versions XXXX01000000. The accession numbers for 16S rRNA sequences are listed in Table S4. Primers used for PCR amplification of internal gene sequences that were used for MLST and MLSA have been published (8, 10). Accession numbers for reference MLST genes (10) are EU075657 to EU076239. The accession numbers of MLST/MLSA sequences generated in this study are listed in Table S4.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://dx.doi.org/10.1128/mSphere.00041-15.

Figure S1, TIF file, 0.2 MB.
Figure S2, TIF file, 0.5 MB.
Figure S3, TIF file, 0.5 MB.
Table S1, XLSX file, 0.04 MB.
Table S2, XLSX file, 0.1 MB.
Table S3, XLSX file, 0.01 MB.
Table S4, XLSX file, 0.03 MB.

ACKNOWLEDGMENTS

We thank Brigitte Rosenberg and Michele Memmer for isolation of genomic DNA and DNA sequencing and Ulrike Klein and Tina Jensen for help during the isolation of bacteria from crude samples and MIC determination. We also are grateful to the Frankfurt Zoo for providing bacterial samples from great apes and to Christophe Boesch for providing access to the wild chimpanzees of Taï National Park. We thank the Ivorian authorities for long-term support, especially the Ministry of the Environment and Forests as well as the Ministry of Research, the directorship of the Taï National Park, and the Swiss Research Center in Abidjan, Ivory Coast.

This work was supported by the Deutsche Forschungsgemeinschaft (grant 1011/13-1 to R.H.). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

FUNDING INFORMATION

This work, including the efforts of Regine Hakenbeck, was funded by Deutsche Forschungsgemeinschaft (DFG) (1011/13-1).

REFERENCES

1. **Mitchell TJ**. 2003. The pathogenesis of streptococcal infections: from tooth decay to meningitis. Nat Rev Microbiol **1:**219–230. http://dx.doi.org/10.1038/nrmicro771.
2. **Shahinas D, Thornton CS, Tamber GS, Arya G, Wong A, Jamieson FB, Ma JH, Alexander DC, Low DE, Pillai DR**. 2013. Comparative genomic analyses of *Streptococcus pseudopneumoniae* provide insight into virulence and commensalism dynamics. PLoS One **8:**e65670. http://dx.doi.org/10.1371/journal.pone.0065670.
3. **Denapaite D, Brückner R, Nuhn M, Reichmann P, Henrich B, Maurer P, Schähle Y, Selbmann P, Zimmermann W, Wambutt R, Hakenbeck R**. 2010. The genome of *Streptococcus mitis* B6—what is a commensal? PLoS One **5:**e9426. http://dx.doi.org/10.1371/journal.pone.0009426.
4. **Tettelin H, Chancey S, Mitchell T, Denapaite D, Schähle Y, Rieger M, Hakenbeck R**. 2015. Genomics, genetic variation, and regions of differences, p 81–107. In Brown J, Hammerschmidt S, Orihuela C (ed), *Streptococcus pneumoniae* molecular mechanisms of host-pathogen interaction. Academic Press, London, United Kingdom.
5. **Kilian M, Riley DR, Jensen A, Brüggemann H, Tettelin H**. 2014. Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. mBio **5:**e01490-14. http://dx.doi.org/10.1128/mBio.01490-14.
6. **Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, Sørensen UB**. 2008. Evolution of *Streptococcus pneumoniae* and its close commensal relatives. PLoS One **3:**e2683. http://dx.doi.org/10.1371/journal.pone.0002683.
7. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG**. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A **95:**3140–3145. http://dx.doi.org/10.1073/pnas.95.6.3140.
8. **Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG**. 2009. Assigning strains to bacterial species via the internet. BMC Biol **7:**3. http://dx.doi.org/10.1186/1741-7007-7-3.
9. **Hanage WP, Fraser C, Spratt BG**. 2005. Fuzzy species among recombinogenic bacteria. BMC Biol **3:**6. http://dx.doi.org/10.1186/1741-7007-3-6.
10. **Chi F, Nolte O, Bergmann C, Ip M, Hakenbeck R**. 2007. Crossing the barrier: evolution and spread of a major class of mosaic *pbp2x* in *S. pneumoniae*, *S. mitis* and *S. oralis*. Int J Med Microbiol **297:**503–512. http://dx.doi.org/10.1016/j.ijmm.2007.02.009.
11. **Tong H, Shang N, Liu L, Wang X, Cai J, Dong X**. 2013. Complete genome sequence of an oral commensal, *Streptococcus oligofermentans* strain AS 1.3089. Genome Announc **1:**e00353-13. http://dx.doi.org/10.1128/genomeA.00353-13.
12. **Woo PC, Tam DM, Leung KW, Lau SK, Teng JL, Wong MK, Yuen KY**. 2002. *Streptococcus sinensis* sp. nov., a novel species isolated from a patient with infective endocarditis. J Clin Microbiol **40:**805–810. http://dx.doi.org/10.1128/JCM.40.3.805-810.2002.
13. **Zbinden A, Mueller NJ, Tarr PE, Spröer C, Keller PM, Bloemberg GV**. 2012. *Streptococcus tigurinus* sp. nov., isolated from blood of patients with endocarditis, meningitis and spondylodiscitis. Int J Syst Evol Microbiol **62:**2941–2945. http://dx.doi.org/10.1099/ijs.0.038299-0.
14. **Gizard Y, Zbinden A, Schrenzel J, François P**. 2013. Whole-genome sequences of *Streptococcus tigurinus* type strain AZ_3a and *S. tigurinus* 1366, a strain causing prosthetic joint infection. Genome Announc **1:**e00210-12. http://dx.doi.org/10.1128/genomeA.00210-12.
15. **Camelo-Castillo A, Benítez-Páez A, Belda-Ferre P, Cabrera-Rubio R, Mira A**. 2014. *Streptococcus dentisani* sp. nov., a novel member of the mitis group. Int J Syst Evol Microbiol **64:**60–65. http://dx.doi.org/10.1099/ijs.0.054098-0.
16. **Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A, Syrjänen R, Spratt BG**. 2005. Using multilocus sequence data to define the pneumococcus. J Bacteriol **187:**6223–6230. http://dx.doi.org/10.1128/JB.187.17.6223-6230.2005.
17. **Reichmann P, Nuhn M, Denapaite D, Brückner R, Henrich B, Maurer P, Rieger M, Klages S, Reinhard R, Hakenbeck R**. 2011. Genome of *Streptococcus oralis* strain uo5. J Bacteriol **193:**2888–2889. http://dx.doi.org/10.1128/JB.00321-11.
18. **Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP**. 2011. Rapid pneumococcal evolution in response to clinical interventions. Science **331:**430–434. http://dx.doi.org/10.1126/science.1198545.
19. **Wyres KL, van Tonder A, Lambertsen LM, Hakenbeck R, Parkhill J, Bentley SD, Brueggemann AB**. 2013. Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated

prior to 1974. BMC Genomics **14:**500. http://dx.doi.org/10.1186/1471-2164-14-500.

20. **Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, Mitchell AM, Quail MA, Andrew PW, Parkhill J, Bentley SD, Mitchell TJ**. 2009. The role of conjugative elements in the evolution of the multi-drug resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81. J Bacteriol **191:**1480–1489. http://dx.doi.org/10.1128/JB.01343-08.
21. **Li X, Alvarez V, Harper WJ, Wang HH**. 2011. Persistent, toxin-antitoxin system-independent, tetracycline resistance-encoding plasmid from a dairy *Enterococcus faecium* isolate. Appl Environ Microbiol **77:**7096–7103. http://dx.doi.org/10.1128/AEM.05168-11.
22. **Blomberg C, Dagerhamn J, Dahlberg S, Browall S, Fernebro J, Albiger B, Morfeldt E, Normark S, Henriques-Normark B**. 2009. Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. J Infect Dis **199:**1032–1042. http://dx.doi.org/10.1086/597205.
23. **Lizcano A, Sanchez CJ, Orihuela CJ**. 2012. A role for glycosylated serine-rich repeat proteins in gram-positive bacterial pathogenesis. Mol Oral Microbiol **27:**257–269. http://dx.doi.org/10.1111/j.2041-1014.2012.00653.x.
24. **Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV**. 2011. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol **9:**467–477. http://dx.doi.org/10.1038/nrmicro2577.
25. **Zähner D, Gandhi AR, Yi H, Stephens DS**. 2011. Mitis group streptococci express variable pilus islet 2 pili. PLoS One **6:**e25124. http://dx.doi.org/10.1371/journal.pone.0025124.
26. **Bagnoli F, Moschioni M, Donati C, Dimitrovska V, Ferlenghi I, Facciotti C, Muzzi A, Giusti F, Emolo C, Sinisi A, Hilleringmann M, Pansegrau W, Censini S, Rappuoli R, Covacci A, Masignani V, Barocchi MA**. 2008. A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. J Bacteriol **190:**5480–5492. http://dx.doi.org/10.1128/JB.00384-08.
27. **Barocchi MA, Ries J, Zogaj X, Hemsley C, Albiger B, Kanth A, Dahlberg S, Fernebro J, Moschioni M, Masignani V, Hultenby K, Taddei AR, Beiter K, Wartha F, von Euler A, Covacci A, Holden DW, Normark S, Rappuoli R, Henriques-Normark B**. 2006. A pneumococcal pilus influences virulence and host inflammatory responses. Proc Natl Acad Sci U S A **103:**2857–2862. http://dx.doi.org/10.1073/pnas.0511017103.
28. **Madhour A, Maurer P, Hakenbeck R**. 2011. Cell surface proteins in *S. pneumoniae*, *S. mitis* and *S. oralis*. Iran J Microbiol **3:**58–67.
29. **Hakenbeck R, Brückner R, Denapaite D, Maurer P**. 2012. Molecular mechanism of beta-lactam resistance in *Streptococcus pneumoniae*. Future Microbiol **7:**395–410. http://dx.doi.org/10.2217/fmb.12.2.
30. **Jensen A, Valdórsson O, Frimodt-Møller N, Hollingshead S, Kilian M**. 2015. Commensal Streptococci Serve as a reservoir for beta-lactam resistance genes in *Streptococcus pneumoniae*. Antimicrob Agents Chemother **59:**3529–3540. http://dx.doi.org/10.1128/AAC.00429-15.
31. **Grebe T, Hakenbeck R**. 1996. Penicillin-binding proteins 2b and 2x of *Streptococcus pneumoniae* are primary resistance determinants for different classes of β-lactam antibiotics. Antimicrob Agents Chemother **40:**829–834.
32. **Lloyd AJ, Gilbey AM, Blewett AM, Wyres KL, van Tonder A, Lambertsen LM, Hakenbeck R, Parkhill J, Bentley SD, Brueggemann AB, El Zoeiby A, Levesque RC, Catherwood AC, Tomasz A, Bugg TD, Roper DI, Dowson CG**. 2008. Characterization of tRNA-dependent peptide bond formation by MurM in the synthesis of *Streptococcus pneumoniae* peptidoglycan. J Biol Chem **283:**6402–6417. http://dx.doi.org/10.1074/jbc.M708105200.
33. **Filipe SR, Severina E, Tomasz A**. 2000. Distribution of the mosaic structured *murM* genes among natural populations of *Streptococcus pneumoniae*. J Bacteriol **182:**6798–6805. http://dx.doi.org/10.1128/JB.182.23.6798-6805.2000.
34. **Todorova K, Maurer P, Rieger M, Becker T, Bui NK, Gray J, Vollmer W, Hakenbeck R**. 2015. Transfer of penicillin resistance from *Streptococcus oralis* to *Streptococcus pneumoniae* identifies *murE* as resistance determinant. Mol Microbiol **97:**866–880. http://dx.doi.org/10.1111/mmi.13070.
35. **Garcia-Bustos JF, Chait BT, Tomasz A**. 1987. Structure of the peptide network of pneumococcal peptidoglycan. J Biol Chem **262:**15400–15405.
36. **Garcia-Bustos J, Tomasz A**. 1990. A biological price of antibiotic resistance: major changes in the peptidoglycan structure of penicillin-resistant pneumococci. Proc Natl Acad Sci U S A **87:**5415–5419. http://dx.doi.org/10.1073/pnas.87.14.5415.
37. **De Pascale G, Lloyd AJ, Schouten JA, Gilbey AM, Roper DI, Dowson CG, Bugg TD**. 2008. Kinetic characterization of lipid II-Ala:alanyl-tRNA ligase (MurN) from *Streptococcus pneumoniae* using semisynthetic aminoacyl-lipid II substrates. J Biol Chem **283:**34571–34579. http://dx.doi.org/10.1074/jbc.M805807200.
38. **Bui NK, Eberhardt A, Vollmer D, Kern T, Bougault C, Tomasz A, Simorre JP, Vollmer W**. 2012. Isolation and analysis of cell wall components from *Streptococcus pneumoniae*. Anal Biochem **421:**657–666. http://dx.doi.org/10.1016/j.ab.2011.11.026.
39. **Percy MG, Gründling A**. 2014. Lipoteichoic acid synthesis and function in gram-positive bacteria. Annu Rev Microbiol **68:**81–100. http://dx.doi.org/10.1146/annurev-micro-091213-112949.
40. **Denapaite D, Brückner R, Hakenbeck R, Vollmer W**. 2012. Biosynthesis of teichoic acids in *Streptococcus pneumoniae* and closely related species: lessons from genomes. Microb Drug Resist **18:**344–358. http://dx.doi.org/10.1089/mdr.2012.0026.
41. **Fischer W, Behr T, Hartmann R, Peter-Katalinić J, Egge H**. 1993. Teichoic acid and lipoteichoic acid of *Streptococcus pneumoniae* possess identical chain structures. A reinvestigation of teichoid acid (C polysaccharide). Eur J Biochem **215:**851–857. http://dx.doi.org/10.1111/j.1432-1033.1993.tb18102.x.
42. **Gisch N, Schwudke D, Thomsen S, Heß N, Hakenbeck R, Denapaite D**. 2015. Lipoteichoic acid of Streptococcus oralis Uo5: a novel biochemical structure comprising an unusual phosphorylcholine substitution pattern compared to Streptococcus pneumoniae. Sci Rep **5:**16718. http://dx.doi.org/10.1038/srep16718.
43. **Hakenbeck R, Madhour A, Denapaite D, Brückner R**. 2009. Versatility of choline metabolism and choline binding proteins in *Streptococcus pneumoniae* and commensal streptococci. FEMS Microbiol Rev **33:**572–586. http://dx.doi.org/10.1111/j.1574-6976.2009.00172.x.
44. **García P, González MP, García E, López R, García JL**. 1999. LytB, a novel pneumococcal murein hydrolase essential for cell separation. Mol Microbiol **31:**1275–1277. http://dx.doi.org/10.1046/j.1365-2958.1999.01238.x.
45. **Kausmally L, Johnsborg O, Lunde M, Knutsen E, Håvarstein LS**. 2005. Choline-binding protein D (CbpD) in *Streptococcus pneumoniae* is essential for competence-induced cell lysis. J Bacteriol **187:**4338–4345. http://dx.doi.org/10.1128/JB.187.13.4338-4345.2005.
46. **Molina R, González A, Stelter M, Pérez-Dorado I, Kahn R, Morales M, Campuzano S, Campillo NE, Mobashery S, García JL, García P, Hermoso JA**. 2009. Crystal structure of CbpF, a bifunctional choline-binding protein and autolysis regulator from *Streptococcus pneumoniae*. EMBO Rep **10:**246–251. http://dx.doi.org/10.1038/embor.2008.245.
47. **Berg KH, Ohnstad HS, Håvarstein LS**. 2012. LytF, a novel competence-regulated murein hydrolase in the genus *Streptococcus*. J Bacteriol **194:**627–635. http://dx.doi.org/10.1128/JB.06273-11.
48. **Biørnstad TJ, Ohnstad HS, Håvarstein LS**. 2012. Deletion of the murein hydrolase CbpD reduces transformation efficiency in *Streptococcus thermophilus*. Microbiology **158:**877–885. http://dx.doi.org/10.1099/mic.0.056150-0.
49. **Guenzi E, Gasc AM, Sicard MA, Hakenbeck R**. 1994. A two-component signal-transducing system is involved in competence and penicillin susceptibility in laboratory mutants of *Streptococcus pneumoniae*. Mol Microbiol **12:**505–515. http://dx.doi.org/10.1111/j.1365-2958.1994.tb01038.x.
50. **Mascher T, Merai M, Balmelle N, de Saizieu A, Hakenbeck R**. 2003. The *Streptococcus pneumoniae cia* regulon: CiaR target sites and transcription profile analysis. J Bacteriol **185:**60–70. http://dx.doi.org/10.1128/JB.185.1.60-70.2003.
51. **Halfmann A, Kovács M, Hakenbeck R, Brückner R**. 2007. Identification of the genes directly controlled by the response regulator CiaR in *Streptococcus pneumoniae*: five out of fifteen promoters drive expression of small noncoding RNAs. Mol Microbiol **66:**110–126. http://dx.doi.org/10.1111/j.1365-2958.2007.05900.x.
52. **Sebert ME, Patel KP, Plotnick M, Weiser JN**. 2005. Pneumococcal HtrA protease mediates inhibition of competence by the CiaRH two-component signaling system. J Bacteriol **187:**3969–3979. http://dx.doi.org/10.1128/JB.187.12.3969-3979.2005.
53. **Schnorpfeil A, Kranz M, Kovács M, Kirsch C, Gartmann J, Brunner I, Bittmann S, Brückner R**. 2013. Target evaluation of the non-coding csRNAs reveals a link of the two-component regulatory system CiaRH to

competence control in *Streptococcus pneumoniae* R6. Mol Microbiol **89:**334–349. http://dx.doi.org/10.1111/mmi.12277.

54. **Dawid S, Sebert ME, Weiser JN**. 2009. Bacteriocin activity of *Streptococcus pneumoniae* is controlled by the serine protease HtrA via post-transcriptional regulation. J Bacteriol **191:**1509–1518. http://dx.doi.org/10.1128/JB.01213-08.
55. **Sebert ME, Palmer LM, Rosenberg M, Weiser JN**. 2002. Microarray-based identification of *htrA*, a *Streptococcus pneumoniae* gene that is regulated by the CiaRH two-component system and contributes to nasopharyngeal colonization. Infect Immun **70:**4059–4067. http://dx.doi.org/10.1128/IAI.70.8.4059-4067.2002.
56. **Ibrahim YM, Kerr AR, McCluskey J, Mitchell TJ**. 2004. Control of virulence by the two-component system CiaR/H is mediated via HtrA, a major virulence factor of *Streptococcus pneumoniae*. J Bacteriol **186:**5258–5266. http://dx.doi.org/10.1128/JB.186.16.5258-5266.2004.
57. **Mascher T, Heintz M, Zähner D, Merai M, Hakenbeck R**. 2006. The CiaRH system of *Streptococcus pneumoniae* prevents lysis during stress induced by treatment with cell wall inhibitors and mutations in *pbp2x* involved in beta-lactam resistance. J Bacteriol **188:**1959–1968. http://dx.doi.org/10.1128/JB.188.5.1959-1968.2006.
58. **Laux A, Sexauer A, Sivaselvarajah D, Kaysen A, Brückner R**. 2015. Control of competence by related non-coding csRNAs in *Streptococcus pneumoniae* R6. Front Genet **6:**246. http://dx.doi.org/10.3389/fgene.2015.00246.
59. **Marx P, Nuhn M, Kovács M, Hakenbeck R, Brückner R**. 2010. Identification of genes for small non-coding RNAs that belong to the regulon of the two-component regulatory system CiaRH in *Streptococcus*. BMC Genomics **11:**661. http://dx.doi.org/10.1186/1471-2164-11-661.
60. **Bergmann S, Hammerschmidt S**. 2006. Versatility of pneumococcal surface proteins. Microbiology **152:**295–303. http://dx.doi.org/10.1099/mic.0.28610-0.
61. **Hilleringmann M, Kohler S, Gámez G, Hammerschmidt S**. 2015. Pneumococcal pili and adhesins, p 309–346. In Brown J, Hammerschmidt S, Orihuela C (ed), *Streptococcus pneumoniae* molecular mechanisms of host-pathogen interaction. Academic Press, London, United Kingdom.
62. **Bergmann S, Wild D, Diekmann O, Frank R, Bracht D, Chhatwal GS, Hammerschmidt S**. 2003. Identification of a novel plasmin(ogen)-binding motif in surface displayed alpha-enolase of *Streptococcus pneumoniae*. Mol Microbiol **49:**411–423. http://dx.doi.org/10.1046/j.1365-2958.2003.03557.x.
63. **Bergmann S, Rohde M, Chhatwal GS, Hammerschmidt S**. 2001. Alpha-enolase of *Streptococcus pneumoniae* is a plasmin(ogen)-binding protein displayed on the bacterial cell surface. Mol Microbiol **40:**1273–1287. http://dx.doi.org/10.1046/j.1365-2958.2001.02448.x.
64. **Irie A, Koyama S, Kozutsumi Y, Kawasaki T, Suzuki A**. 1998. The molecular basis for the absence of N-glycolylneuraminic acid in humans. J Biol Chem **273:**15866–15871. http://dx.doi.org/10.1074/jbc.273.25.15866.
65. **Chou HH, Takematsu H, Diaz S, Iber J, Nickerson E, Wright KL, Muchmore EA, Nelson DL, Warren ST, Varki A**. 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc Natl Acad Sci U S A **95:**11751–11756. http://dx.doi.org/10.1073/pnas.95.20.11751.
66. **Ogunniyi AD, Paton JC**. 2015. Vaccine potential of pneumococcal proteins, p 59–78. In Brown J, Hammerschmidt S, Orihuela C (ed), *Streptococcus pneumoniae* molecular mechanisms of host-pathogen interaction. Academic Press, London, United Kingdom.
67. **Hammerschmidt S, Bethe G, Remane PH, Chhatwal GS**. 1999. Identification of pneumococcal surface protein A as a lactoferrin-binding protein of *Streptococcus pneumoniae*. Infect Immun **67:**1683–1687.
68. **Tu AH, Fulgham RL, McCrory MA, Briles DE, Szalai AJ**. 1999. Pneumococcal surface protein A inhibits complement activation by *Streptococcus pneumoniae*. Infect Immun **67:**4720–4724.
69. **Hollingshead SK, Becker R, Briles DE**. 2000. Diversity of PspA: mosaic genes and evidence for past recombination in *Streptococcus pneumoniae*. Infect Immun **68:**5889–5900. http://dx.doi.org/10.1128/IAI.68.10.5889-5900.2000.
70. **Glover DT, Hollingshead SK, Briles DE**. 2008. *Streptococcus pneumoniae* surface protein PcpA elicits protection against lung infection and fatal sepsis. Infect Immun **76:**2767–2776. http://dx.doi.org/10.1128/IAI.01126-07.
71. **Hammerschmidt S, Talay SR, Brandtzaeg P, Chhatwal GS**. 1997. SpsA, a novel pneumococcal surface protein with specific binding to secretory immunoglobulin A and secretory component. Mol Microbiol **25:**1113–1124. http://dx.doi.org/10.1046/j.1365-2958.1997.5391899.x.
72. **Voss S, Hallström T, Saleh M, Burchhardt G, Pribyl T, Singh B, Riesbeck K, Zipfel PF, Hammerschmidt S**. 2013. The choline-binding protein PspC of *Streptococcus pneumoniae* interacts with the C-terminal heparin-binding domain of vitronectin. J Biol Chem **288:**15614–15627. http://dx.doi.org/10.1074/jbc.M112.443507.
73. **Janulczyk R, Iannelli F, Sjoholm AG, Pozzi G, Bjorck L**. 2000. Hic, a novel surface protein of *Streptococcus pneumoniae* that interferes with complement function. J Biol Chem **275:**37257–37263. http://dx.doi.org/10.1074/jbc.M004572200.
74. **Iannelli F, Oggioni MR, Pozzi G**. 2002. Allelic variation in the highly polymorphic locus *pspC* of *Streptococcus pneumoniae*. Gen 284:63–71. http://dx.doi.org/10.1016/S0378-1119(01)00896-4.
75. **Briese T, Hakenbeck R**. 1985. Interaction of the pneumococcal amidase with lipoteichoic acid and choline. Eur J Biochem **146:**417–427. http://dx.doi.org/10.1111/j.1432-1033.1985.tb08668.x.
76. **Giudicelli S, Tomasz A**. 1984. Inhibition of the in vitro and in vivo activity of the pneumococcal autolytic enzyme—by choline and phosphorylcholine, p 207–212. In Nombela C (ed), Microbial cell wall synthesis and autolysis. Elsevier Science Publishers, Amsterdam, Netherlands.
77. **Mellroth P, Daniels R, Eberhardt A, Rönnlund D, Blom H, Widengren J, Normark S, Henriques-Normark B**. 2012. LytA, major autolysin of *Streptococcus pneumoniae*, requires access to nascent peptidoglycan. J Biol Chem **287:**11018–11029. http://dx.doi.org/10.1074/jbc.M111.318584.
78. **Mellroth P, Sandalova T, Kikhney A, Vilaplana F, Hesek D, Lee M, Mobashery S, Normark S, Svergun D, Henriques-Normark B, Achour A**. 2014. Structural and functional insights into peptidoglycan access for the lytic amidase LytA of *Streptococcus pneumoniae*. mBio **5:**e01120-13. http://dx.doi.org/10.1128/mBio.01120-13.
79. **Claverys JP, Håvarstein LS**. 2007. Cannibalism and fratricide: mechanisms and raisons d'etre. Nat Rev Microbiol **5:**219–229. http://dx.doi.org/10.1038/nrmicro1613.
80. **Berry AM, Lock RA, Hansman D, Paton JC**. 1989. Contribution of autolysin to virulence of *Streptococcus pneumoniae*. Infect Immun **57:**2324–2330.
81. **Whatmore AM, Dowson CG**. 1999. The autolysin-encoding gene (*lytA*) of *Streptococcus pneumoniae* displays restricted allelic variation despite localized recombination events with genes of pneumococcal bacteriophage encoding cell wall lytic enzymes. Infect Immun **67:**4551–4556.
82. **Kearns AM, Wheeler J, Freeman R, Seiders PR, Perry J, Whatmore AM, Dowson CG**. 2000. Pneumolysin detection identifies atypical isolates of *Streptococcus pneumoniae*. J Clin Microbiol **38:**1309–1310.
83. **Jefferies J, Nieminen L, Kirkham LA, Johnston C, Smith A, Mitchell TJ**. 2007. Identification of a secreted cholesterol-dependent cytolysin (mitilysin) from *Streptococcus mitis*. J Bacteriol **189:**627–632. http://dx.doi.org/10.1128/JB.01092-06.
84. **Miyanohara M, Imai S, Okamoto M, Saito W, Nomura Y, Momoi Y, Tomonaga M, Hanada N**. 2013. Distribution of *Streptococcus troglodytae* and *Streptococcus dentirousetti* in chimpanzee oral cavities. Microbiol Immunol **57:**359–365. http://dx.doi.org/10.1111/1348-0421.12047.
85. **Calvignac-Spencer S, Leendertz SA, Gillespie TR, Leendertz FH**. 2012. Wild great apes as sentinels and sources of infectious disease. Clin Microbiol Infect **18:**521–527. http://dx.doi.org/10.1111/j.1469-0691.2012.03816.x.
86. **Schaumburg F, Mugisha L, Kappeller P, Fichtel C, Köck R, Köndgen S, Becker K, Boesch C, Peters G, Leendertz F**. 2013. Evaluation of non-invasive biological samples to monitor *Staphylococcus aureus* colonization in great apes and lemurs. PLoS One **8:**e78046. http://dx.doi.org/10.1371/journal.pone.0078046.
87. **Zbinden A, Mueller NJ, Tarr PE, Eich G, Schulthess B, Bahlmann AS, Keller PM, Bloemberg GV**. 2012. *Streptococcus tigurinus*, a novel member of the *Streptococcus mitis* group, causes invasive infections. J Clin Microbiol **50:**2969–2973. http://dx.doi.org/10.1128/JCM.00849-12.
88. **Hakenbeck R, Balmelle N, Weber B, Gardès C, Keck W, de Saizieu A**. 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies variation of *Streptococcus pneumoniae*. Infect Immun **69:**2477–2486. http://dx.doi.org/10.1128/IAI.69.4.2477-2486.2001.
89. **Weber B, Ehlert K, Diehl A, Reichmann P, Labischinski H, Hakenbeck R**. 2000. The *fib* locus in *Streptococcus pneumoniae* is required for peptidoglycan crosslinking and PBP-mediated beta-lactam resistance. FEMS Microbiol Lett **188:**81–85. http://dx.doi.org/10.1016/S0378-1097(00)00214-7.

90. **Oggioni MR, Claverys J-P**. 1999. Repeated extragenic sequences in procaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. Microbiology **145:**2647–2653. http://dx.doi.org/10.1099/00221287-145-10-2647.
91. **Springer A, Razafimanantsoa L, Fichtel C, Kappeler PM**. 2015. Comparison of three short-term immobilization regimes in wild verreaux's sifakas (*Prophitecus verreauxi*): ketamine-xylazine, ketamine-xylazine-atropine, and tiletamine-zolazepam. J Zoo Wildl Med **46:**482–490. http://dx.doi.org/10.1638/2014-0154.1.
92. **O'Brien KL, Bronsdon MA, Dagan R, Yagupsky P, Janco J, Elliott J, Whitney CG, Yang Y-H, Robinson L-GE, Schwartz B, Carlone GM**. 2001. Evaluation of a medium (STGG) for transport and optimal recovery of Streptococcus pneumoniae from nasopharyngeal secretions collected during field studies. J Clin Microbiol **39:**1021–1024. http://dx.doi.org/10.1128/JCM.39.3.1021-1024.2001.
93. **Lacks S, Hotchkiss RD**. 1960. A study of the genetic material determining an enzyme activity in pneumococcus. Biochim Biophys Acta **39:**508–517. http://dx.doi.org/10.1016/0006-3002(60)90205-5.
94. **Alloing G, Granadel C, Morrison DA, Claverys J-P**. 1996. Competence pheromone, oligopeptide permease, and induction of competence in *Streptococcus pneumoniae*. Mol Microbiol **21:**471–478. http://dx.doi.org/10.1111/j.1365-2958.1996.tb02556.x.
95. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O**. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9:**75. http://dx.doi.org/10.1186/1471-2164-9-75.
96. **Jones CE, Baumann U, Brown AL**. 2005. Automated methods of predicting the function of biological sequences using GO and BLAST. BMC Bioinformatics **6:**272. http://dx.doi.org/10.1186/1471-2105-6-272.
97. **Sibold C, Henrichsen J, König A, Martin C, Chalkley L, Hakenbeck R**. 1994. Mosaic *pbpX* genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from *pbpX* genes of a penicillin-sensitive *Streptococcus oralis*. Mol Microbiol **12:**1013–1023. http://dx.doi.org/10.1111/j.1365-2958.1994.tb01089.x.
98. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S**. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol **30:**2725–2729. http://dx.doi.org/10.1093/molbev/mst197.

2.4 Transfer of penicillin resistance from *Streptococcus oralis* to *Streptococcus pneumoniae* identifies *murE* as resistance determinant

Katya Todorova, Patrick Maurer, **Martin Rieger**, Tina Becker, Nhat Khai Bui, Joe Gray, Waldemar Vollmer and Regine Hakenbeck. Mol Microbiol. 2015 Sep;97(5):866-80. doi: 10.1111/mmi.13070. Epub 2015 Jun 19.

Summary:

In *Streptococcus pneumoniae*, penicillin binding proteins (PBP), as well as MurM and MurN, are important enzymes involved in peptidoglycan (PG) synthesis. Resistance against the beta-lactam antibiotics in clinical isolates is mainly due to alterations in the genes *pbp2x*, *pbp2b*, *pbp1a*, and *murM*. Transformation experiments with DNA from the high-level beta-lactam resistant *S. oralis* strain Uo5 into the recipient *S. pneumoniae* R6 strain were performed to identify genes involved in the high resistance level of the donor strain. The genome sequence of a high-level resistant transformant PCP indicated that the gene *murE*, previously not associated with the evolution of penicillin resistance, contributes to this phenotype, an assumption which could be confirmed experimentally. MurE adds a lysine residue to the PG precursor. Like the three penicillin-binding protein genes and *murM*, *murE* is a mosaic gene in *S. oralis* Uo5, and thus has apparently been imported into this strain. MurE genes with sequence blocks identical to *murE* of *S. oralis* Uo5 were recognized in some *S. pneumoniae* and *S. mitis* strains as well. Unlike *S. pneumoniae*, *S. oralis* Uo5 does not contain MurN, which is reflected in its distinct PG biochemistry. The study added MurE as an important resistance determinant, underlining the importance of non-PBP genes in the development of penicillin resistance.

Own contribution to the paper:

Assembly of sequence reads, final generation of genome sequences from sequence reads and contigs including annotation of three *S. pneumoniae* transformants (PCP-7, PCP-C6 and PCP-CCO). Comparative analysis of the three genomes and with *S. pneumoniae* R6 and *S. oralis* Uo5

including manual retrieval of SNVs and transferred sequence fragments as described in chapter 3.4.

Molecular Microbiology (2015) ■ doi:10.1111/mmi.13070

Transfer of penicillin resistance from *Streptococcus oralis* to *Streptococcus pneumoniae* identifies *murE* as resistance determinant

Katya Todorova,[1] Patrick Maurer,[1†] Martin Rieger,[1] Tina Becker,[1] Nhat Khai Bui,[2‡] Joe Gray,[3] Waldemar Vollmer[2] and Regine Hakenbeck[1*]
[1]*Department of Microbiology, University of Kaiserslautern, Kaiserslautern, Germany.*
[2]*Centre for Bacterial Cell Biology, Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4AX, UK.*
[3]*Institute for Cell and Molecular Biosciences, Pinnacle Laboratory, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.*

Summary

Beta-lactam resistant clinical isolates of *Streptococcus pneumoniae* contain altered penicillin-binding protein (PBP) genes and occasionally an altered *murM*, presumably products of interspecies gene transfer. MurM and MurN are responsible for the synthesis of branched lipid II, substrate for the PBP catalyzed transpeptidation reaction. Here we used the high-level beta-lactam resistant *S. oralis* Uo5 as donor in transformation experiments with the sensitive laboratory strain *S. pneumoniae* R6 as recipient. Surprisingly, piperacillin-resistant transformants contained no alterations in PBP genes but carried $murE_{Uo5}$ encoding the UDP-N-acetylmuramyl tripeptide synthetase. Codons 83–183 of $murE_{Uo5}$ were sufficient to confer the resistance phenotype. Moreover, the promoter of $murE_{Uo5}$, which drives a twofold higher expression compared to that of *S. pneumoniae* R6, could also confer increased resistance. Multiple independent transformations produced *S. pneumoniae* R6 derivatives containing $murE_{Uo5}$, $pbp2x_{Uo5}$, $pbp1a_{Uo5}$ and $pbp2b_{Uo5}$, but not $murM_{Uo5}$ sequences; however, the resistance level of the donor strain could not be reached. *S. oralis* Uo5 harbors an unusual *murM*, and *murN* is absent. Accordingly, the peptidoglycan of *S. oralis* Uo5 contained interpeptide bridges with one L-Ala residue only. The data suggest that resistance in *S. oralis* Uo5 is based on a complex interplay of distinct PBPs and other enzymes involved in peptidoglycan biosynthesis.

Accepted 21 May, 2015. *For correspondence. E-mail hakenb@rhrk.uni-kl.de; Tel. (+49) 631 205 2353; Fax (+49) 631 205 3799. Present addresses: †University of Applied Sciences, Goebenstrasse 40, Saarbrücken, Germany; ‡Helmholtz Centre for Infection Research, Helmholtz Institute for Pharmaceutical Research, Saarland University, 66123 Saarbrücken, Germany.

Introduction

Development of beta-lactam resistance in *Streptococcus pneumoniae* has become a paradigm of evolutionary processes in the antibiotic era. Since the late 1970s, penicillin-resistant *S. pneumoniae* (PRSP) has increased worldwide to represent now over 50% of all isolates in many countries, and the resistance level has risen from the minimal inhibitory concentration (MIC) values of around 0.01 μg ml^{-1} for benzylpenicillin – typical of sensitive isolates – to over 100-fold in PRSP (for review, see Henriques-Normark, 2007). The wide range of resistance levels suggests a complicated multistep process that underlies this phenomenon.

Penicillin resistance is mainly due to alterations in the target enzymes for beta-lactam antibiotics, the penicillin-binding proteins (PBPs). PBPs act during late steps of the biosynthesis of the essential cell wall peptidoglycan. They are modular enzymes containing a transpeptidase domain responsible for cross-linking the peptide subunits, the crucial penicillin-sensitive step. PBPs are grouped into three main classes according to their primary sequence, domain structure and enzymatic activities: class A PBPs contain an N-terminal glycosyltransferase domain, which in class B PBPs is replaced by an N-terminal domain of unknown function; class C PBPs act as DD-carboxypeptidase and/or endopeptidase (Goffin and Ghuysen, 1998; Sauvage *et al.*, 2008). *S. pneumoniae* contains six PBPs: the three class A PBPs 1a, 1b and 2a, the essential class B PBPs 2x and 2b, and the DD-carboxypeptidase PBP3. Altered PBP2x, 2b and 1a are the main players during the development of beta-lactam resistance (for review, see Hakenbeck *et al.*, 2012). Beta-lactams are bound covalently to the active site serine residue within the transpeptidase domains of PBPs. Resistant strains harbor mutations that result in a low

affinity for beta-lactam antibiotics, apparently enabling enzymes to function in cell wall synthesis in the presence of beta-lactams. In some resistant clinical isolates, a low affinity PBP2a contributes to resistance (Smith *et al.*, 2005), whereas resistant laboratory mutants can have sequence alterations in all PBPs (Hakenbeck *et al.*, 2012).

PBP2x and PBP2b are primary resistance determinants, i.e. they are capable of conferring low or intermediate levels of resistance when transformed into a penicillin-sensitive strain. However, high resistance levels can only be achieved upon additional introduction of an altered PBP1a. PBP2x plus PBP1a suffice to confer high resistance levels to cefotaxime and related antibiotics (Muñóz *et al.*, 1992), as PBP2b does not interact with these beta-lactams (Hakenbeck *et al.*, 1987). Consequently, mutations in PBP2b can only be selected with penicillins, whereas cefotaxime favors the selection of PBP2x mutations.

Non-PBP genes are also involved in the resistance phenotype in some PRSP. Some penicillin-resistant strains express an altered, mosaic MurM gene (Filipe *et al.*, 2001; Cafini *et al.*, 2005; del Campo *et al.*, 2006). MurM adds an L-Ala or L-Ser to the ε-amino group of the L-Lys residue of lipid II, and MurN adds another L-Ala residue. The resulting branches are used as acceptor substrate for the transpeptidation reaction of PBPs, becoming the interpeptide bridges in mature peptidoglycan (Lloyd *et al.*, 2008). Resistant isolates with an altered *murM* produce much higher amounts of peptide branches (L-Ala-L-Ala or L-Ser-L-Ala) than sensitive and other resistant clones (Severin and Tomasz, 1996). Surprisingly, deletion of *murM* results in a complete breakdown of resistance in all resistant strains (Filipe and Tomasz, 2000; Weber *et al.*, 2000), a phenomenon that is not well understood. Moreover, mutations in the histidine protein kinase CiaH, which occur frequently in laboratory mutants but are rare in resistant clinical isolates (Müller *et al.*, 2011, and references within), can contribute to resistance.

PBP genes and occasionally *murM* in PRSP have a mosaic structure where sequence blocks are replaced by highly altered sequences as a result of interspecies gene transfer events (Dowson *et al.*, 1989; Laible *et al.*, 1991; Martin *et al.*, 1992; Filipe *et al.*, 2001). Sequences related to those of mosaic blocks of the resistant PBP genes have been identified in sensitive *S. mitis*, the closest relatives of *S. pneumoniae* (Dowson *et al.*, 1993; Sibold *et al.*, 1994), indicating that 'resistant' PBP genes have evolved in *S. mitis* prior to transfer into other species. In fact, using DNA of resistant *S. mitis* and *S. oralis* as donor, transformants of *S. pneumoniae* can be obtained in the laboratory expressing mosaic PBP variants, confirming that interspecies gene transfer can occur (Reichmann *et al.*, 1997; Hakenbeck *et al.*, 1998). Indeed, genome-wide analysis of resistant derivatives of the unencapsulated sensitive laboratory strain *S. pneumoniae* R6 obtained after several consecutive transformations with DNA from the high level resistant strain *S. mitis* B6 documented multiple recombination events scattered throughout the genome that included all PBP genes (except the PBP3 gene) and *murM* (Sauerbier *et al.*, 2012). Sequence comparison revealed that the donor strain *S. mitis* B6 already contained PBP genes and *murM* of apparent mosaic structure, suggesting that also in this species high resistance levels are the result of multiple intra- and possibly interspecies DNA transfer events.

We have now used the high level penicillin and multiple antibiotic resistant *S. oralis* Uo5 strain as donor to perform multiple transformations with chromosomal DNA and PCR amplified fragments using *S. pneumoniae* R6 as recipient. *S. oralis* Uo5 was used as its genome sequence is available (Reichmann *et al.*, 2011), enabling the identification of known penicillin resistance determinants and thus facilitating genetic approaches. Our main task was to see which resistance level can be achieved and whether non-PBP genes are involved as well.

Results

Transformation of S. pneumoniae *with chromosomal* S. oralis *Uo5 DNA – how high can we get?*

Streptococcus oralis Uo5 is a high penicillin and cefotaxime resistant isolate (Reichmann *et al.*, 2011) containing low affinity PBP2x, PBP2b and PBP1a. *S. oralis* Uo5 DNA has been used in transformation experiments previously with the sensitive *S. pneumoniae* R6 laboratory strain as recipient (Reichmann *et al.*, 1997). With cefotaxime as the selective antibiotic, only transformants containing $pbp2x_{Uo5}$ were obtained. We now used other beta-lactams in addition to cefotaxime for the selection of beta-lactam resistance. First, we wanted to test whether $pbp2b_{Uo5}$, the other primary resistance determinant, can also be introduced into *S. pneumoniae* R6. This required the selection with oxacillin or piperacillin because cefotaxime does not select for low affinity PBP2b (Hakenbeck *et al.*, 1987). Second, we used cefotaxime to select for changes in *pbp2x* (Grebe and Hakenbeck, 1996) and, in a further round, for an altered *pbp1a* (Muñóz *et al.*, 1992). Finally, multiple transformations were performed with chromosomal Uo5 DNA to see which resistance level can be transferred. Transformants were examined for PBP changes and for potential other resistance determinants. In each step several narrow concentrations of the selective beta-lactam were used at and above the MIC of the recipient, and always the transformant with the highest resistance level was chosen for the next selection step (Fig. 1A; see *Experimental procedures* for detail).

With all three beta-lactams, transformants were obtained in one transformation step. All cefotaxime-

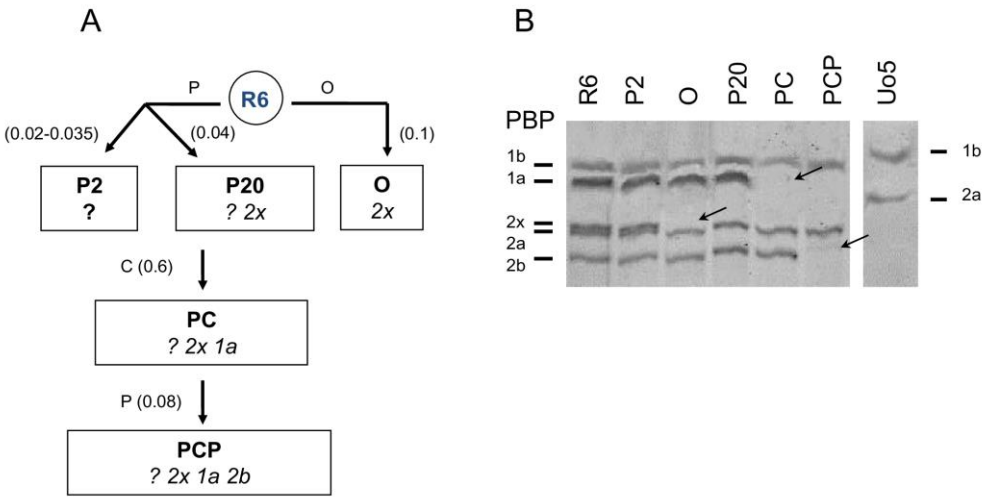


Fig. 1. *S. pneumoniae* R6 transformants obtained with chromosomal DNA of *S. oralis* Uo5.
A. Schematic representation of the selection steps. O (oxacillin), P (piperacillin) and C (cefotaxime) indicate the beta-lactams used for selection. Transformants are named according to the order of selective beta-lactams used during three transformation steps. Genes of the resistance determinants that are present in the transformants are indicated by 2x (*pbp2x*), 1a (*pbp1a*) and 2b (*pbp2b*). The selective concentrations are indicated in brackets. The question mark (?) indicates a non-PBP resistance determinant.
B. PBP-profiles of the transformants. PBPs of cell lysates were labeled with Bocillin™FL and separated by SDS-PAGE followed by fluorography. Arrows depict the low affinity PBPs in the transformants. PBPs of the recipient *S. pneumoniae* R6 are indicated on the left.

resistant transformants (C) contained a low affinity PBP2x (not shown) as did all oxacillin-resistant transformants (O; Fig. 1B). Curiously, 19 transformants obtained with piperacillin (P) at selective concentrations between 0.02 and 0.035 μg ml^{-1} contained no PBP changes (P2 is shown as an example in Fig. 1B), whereas P20, the only transformant obtained at 0.04 μg ml^{-1} piperacillin, again contained a low affinity PBP2x (Fig. 1B). These results are consistent with the previously reported selection of PBP2x mutations with penicillins (Zerfaß *et al.*, 2009). However, no transformants with a low affinity PBP2b were obtained.

Selected transformants O and C expressing a low affinity PBP2x displayed the same resistance profile for the three beta-lactams (only the values for the transformant O are shown in Fig. 2). Compared with these, the transformant P20 had between twofold and fourfold higher MIC for all three antibiotics and therefore was used in the next transformation step as recipient. Transformants containing a low affinity PBP1a could be selected with 0.6 μg ml^{-1} cefotaxime (PC), and a low affinity PBP2b could be introduced in a third transformation using piperacillin (0.08 μg ml^{-1}) resulting in the transformant PCP. The further selection of higher resistant transformants with cefotaxime and oxacillin proved difficult. The number of colonies on the selective plates was low, suggesting that the selection of point mutations rather than gene transfer has occurred, and this was confirmed by further experiments (own unpubl. data). Thus, the maximum of resistance that could be transformed with chromosomal DNA of *S. oralis* Uo5 to *S. pneumoniae* R6 was well below the level of the donor *S. oralis* Uo5, which expresses an approximately 20-fold higher MIC for piperacillin, over threefold for cefotaxime, and more than twofold for oxacillin compared with PCP (Fig. 2).

The presence of *S. oralis* Uo5 PBP sequences was verified by DNA sequencing: P20 contained $pbp2x_{Uo5}$ sequence from nucleotide (nt) 853 to nt 2980 into the downstream gene *mraY*, in PC the entire $pbp1a_{Uo5}$ was present plus parts of the flanking genes *recU* and spr0326 (region from nt −156 to +2610) and $pbp2b_{Uo5}$ in PCP included upstream regions and parts of the transpeptidase domain (nt −1177 to 1350). All three PBPs differed widely from the sequences of *S. pneumoniae* R6 PBPs with changes affecting more than 15% of amino acids (Figs S1–S3).

These data suggest that *S. oralis* Uo5 PBP sequences can confer certain level of beta-lactam resistance in *S. pneumoniae*, but other genes important for high level resistance must be present in *S. oralis* Uo5 and these cannot be readily transferred to *S. pneumoniae*. Moreo-

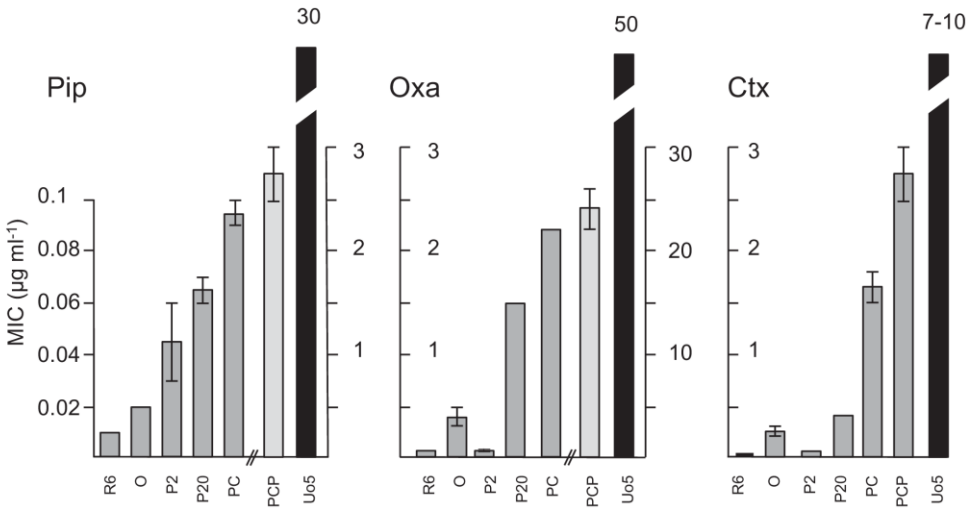


Fig. 2. Resistance pattern of *S. pneumoniae* R6 transformants obtained with chromosomal DNA of *S. oralis* Uo5. The three beta-lactams oxacillin (Oxa), piperacillin (Pip) and cefotaxime (Ctx) were used. Mean values of at least three independent experiments are shown. Bars indicate standard deviation. Note the different scale valid for PCP transformants as indicated by the different shading. The MIC values of the donor strain *S. oralis* Uo5 are indicated above the black bar.

ver, it is likely that a non-PBP resistance determinant selectable with piperacillin has been transferred into the P transformants.

PBP2x$_{Uo5}$ and PBP2b$_{Uo5}$ genes are not the only resistance determinants

In order to verify that a non-PBP gene conferring beta-lactam resistance is involved in the resistance phenotype of P and P20 transformants, PCR fragments of *S. oralis* Uo5 PBP genes including approximately 1 kb flanking regions were used to transform *S. pneumoniae* R6. Transformants obtained with PCR fragments are designated with *. Oxacillin was used to select for transfer of *pbp2x*$_{Uo5}$. Oxacillin-resistant transformants O*2x containing a low affinity PBP2x were readily obtained, and they all showed significantly increased MIC values to oxacillin and cefotaxime and slightly increased resistance to piperacillin, similar to the O and C transformants obtained with chromosomal DNA (Fig. 3). However, none of the transformants reached the MIC levels expressed in P20.

Piperacillin at a concentration just above the MIC of the recipient (0.02 µg ml^{-1}) was chosen to select for the introduction of PBP2b$_{Uo5}$ into R6. Only one out of 10 transformants examined contained a low affinity PBP2b (P*2b). The MIC was similar to that observed in the P mutants described above (Fig. 3). Curiously, only the *pbp2b*$_{Uo5}$ region around the codon changing Thr446 to Ala was introduced into the two transformants P*2b and OP*2x2b: P*2b contained the region between nt 1331–1350 resulting in the single T446A exchange, and OP*2xb had nt 1261–1350 or four aa alterations in addition to T446A (Fig. S3). None of the transformants contained *S. oralis* sequences after codon 450. Higher resistance levels to penicillins than in P20 were obtained only after transforming the PBP2b gene into O*2x.

The MurE gene of S. oralis *Uo5 is altered in piperacillin-resistant transformants*

To identify genes altered in P20 in addition to *pbp2x* comparative genomic hybridization on a *S. pneumoniae* R6 oligonucleotide microarray was performed using DNA of *S. pneumoniae* R6 versus P20 (not shown). Genes located in three regions of the genomes displayed low hybridization signals indicating the presence of *S. oralis* DNA. The first region included *mraY* (*spr0305*), which is located downstream of *pbp2x* (*pbp2x* was not detected since the oligonucleotide of the microarray matched also *pbp2x*$_{Uo5}$ sequences), the second corresponded to *xerC* (*spr1046*), a recombinase gene that is unlikely to be involved in resistance, and the third region included *murE* (*spr1384*) encoding the UDP-N-acetylmuramyl tripeptide synthetase and hypothetical genes located downstream of *murE*.

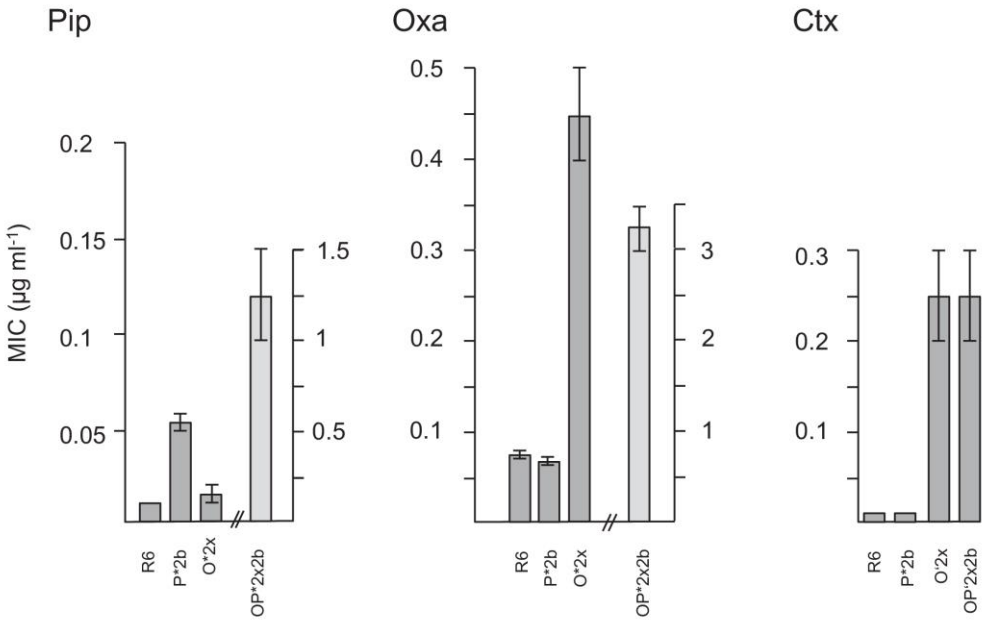


Fig. 3. Resistance pattern of *S. pneumoniae* R6 transformants containing PBP2x$_{Uo5}$, PBP2b$_{Uo5}$ or both PBPs. The three beta-lactams oxacillin (Oxa), piperacillin (Pip) and cefotaxime (Ctx) were used. Note the two different scales for MIC Oxa and Pip values used for the double transformants.

Next, the genome of PCP was sequenced to identify recombined sequences of *S. oralis* Uo5. In addition to the three PBP genes encoding low affinity PBP2x, 1a and 2b, the only region that drew our attention again included *murE* (*spr1384* encoding the UDP-N-acetylmuramyl tripeptide synthetase) and affected the flanking genes *spr1383*, *spr1385* and *spr1386*. Another five recombination events were detected, but their putative gene products did not indicate a role in resistance (Fig. 4 and Table S1). No changes in other PBP genes nor in *murM* were observed.

Because *murE* has been shown to be involved in the resistance phenotype of *Staphylococcus aureus* (Gardete *et al.*, 2004), the gene locus was sequenced in P20 (Fig. 5) and in three of the resistant P transformants containing no apparent PBP changes. All these transformants contained the sequence of the corresponding *S. oralis* DNA upstream of *murE* or in various parts of the coding region. P20 and P15 contained the entire *murE*$_{Uo5}$, P5 only the first 750 nt and P2 only the first 240 bp of the coding regions.

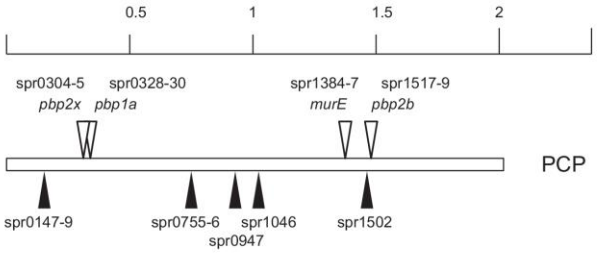


Fig. 4. Recombination sites in the transformant PCP. White arrows mark the position of the resistance determinants *pbp2x*, *pbp1a*, *murE* and *pbp2b*; other recombination sites are marked by black arrows. The *S. pneumoniae* R6 genes affected are indicated.

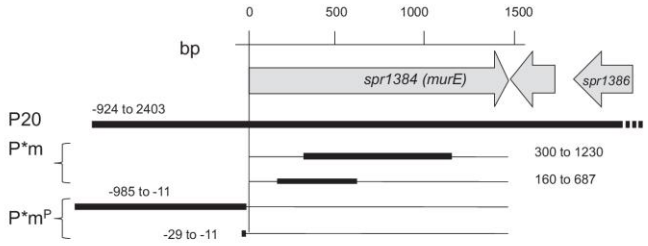


Fig. 5. Schematic representation of the mosaic MurE gene structure in *S. pneumoniae* transformants obtained with *S. oralis* Uo5 DNA. The transformant P20 was obtained with chromosomal DNA. The sequence of two transformants each P*m (obtained with a PCR fragment covering $murE_{Uo5}$) and P*m^{P} (obtained with a PCR fragment covering the promoter and upstream regions) are shown. +1 in the scale corresponds to the start of the MurE gene. The black bars indicate the recombined regions.

murE$_{Uo5}$ *confers increase in beta-lactam resistance*

To confirm that *murE*$_{Uo5}$ confers a resistance phenotype in R6, PCR amplified fragments of *murE*$_{P20}$ containing the entire gene were used in transformations with R6 and O*2x as recipients, using piperacillin for selection at concentrations at and slightly above the MIC of the recipient. Transformants P**murE* and O*2x*murE* were readily obtained that displayed MIC values similar to P2 respectively P20 (see Fig. 2). DNA sequencing confirmed the presence of *murE*$_{Uo5}$: O*2x*murE* contained the entire coding sequence, whereas two examined P**murE* transformants contained internal sequences overlapping between nt 300 and 687 (Fig. 5). This sequence block results in protein changes between aa 113 and 183 corresponding to 15 aa differences compared with MurE$_{R6}$ (Fig. 6A).

Since these sequences were not present in P2, a PCR fragment obtained from *murE* of the P2 transformant, covering only the first 240 nt plus upstream region, was used in a second set of transformation experiments with R6 as recipient. Selective conditions were as described above. Transformants were readily obtained with MIC values similar to P2. Two transformants were sequenced, and both contained only upstream sequences that included at least the promoter region (Fig. 5). The promoter of *murE*$_{Uo5}$ is highly altered compared with that of *S. pneumoniae* R6 with one change C to T being located within the −10 box and another seven in the region between the −10 and −35 boxes (Fig. 6B). This strongly suggests that the promoter region also confers the resistance phenotype.

The promoter of murE$_{Uo5}$ *drives higher expression compared with* murE$_{R6}$ *and confers β-lactam resistance*

If the *murE*$_{Uo5}$ promoter region is involved in resistance, it could be due to altered expression levels of the MurE gene. To test this assumption, the promoter region of *murE*$_{Uo5}$ and for comparison that of *murE*$_{R6}$ were cloned into the pPP2 promoter probe plasmid containing a promoterless β-galactosidase gene *lacZ* (Halfmann *et al.*,

A

```
          1111111111111111111111111111111111111111111111111111111111111111111111111111
          1111111222222222233333333334444444444555555555566666666667777777777888888
          3456789012345678901234567890123456789012345678901234567890123456789012345678
So Uo5    ILAFTGTKGKTTAAYFAYHILSQRYPTALLSTMNTTLDGETFFKSSFSTPENIDLFDMMAQAVQNGRTHLV
So 49296  ..........................................................K.......
Sm B6     ....................................K.....................K.......
Sp 670    ....................................T.....................K...S...
Sp R6     L...........T....N....GHRP.M.............ALT...S.......N.....D....I
```

B

```
                                        -35                 -10          murE
So Uo5    ATCTTCCATTATAAACTTTTCCTTGTTACTTGTAAAATTCGGGCTTGAAGTTTATAATAGAAAGTATG
So 49296  ..........................................A.A.......................
Sm B6     ..........................C.....................................
Sp R6     ..................................T..AC.A...TCG.....C.............
```

Fig. 6. A. Protein sequences of MurE variants related to the MurE$_{Uo5}$ region aa 113–183 common in R6 transformants to those in other *Streptococcus* spp. The positions of the amino acids are indicated vertically in the three rows on top. So, *S. oralis*; Sm, *S. mitis*; Sp, *S. pneumoniae*. Protein sequences identical to *S. pneumoniae* 670 are also present in *S. pneumoniae* TCH8431/19A, A026, Taiwan19F-14, ST556. B. Comparison of the *murE*$_{Uo5}$ promoter region with those of other *Streptococcus* spp. The −35 and −10 region are underlined.

Table 1. R6 derivatives with an ectopic copy of *murE* carrying various promoter regions and coding sequences.[a]

Strain[b]	Genotype	MIC (μg ml^{-1})[a] Pip	Oxa	Ctx
O*xUU	*pbp2x* Δ*murE*$_{R6}$P$_{Uo5}$*murE*$_{Uo5}$	0.07–0.08	1–1.5	0.4
O*xRU	*pbp2x* Δ*murE*$_{R6}$P$_{R6}$*murE*$_{Uo5}$	0.07–0.08	1–1.5	0.4
O*xUR	*pbp2x* Δ*murE*$_{R6}$P$_{Uo5}$*murE*$_{R6}$	0.07–0.08	1–1.5	0.4
O*xRR	*pbp2x* Δ*murE*$_{R6}$P$_{R6}$*murE*$_{R6}$	0.02–0.03	0.4–0.5	0.3

a. MIC values were taken from at least three independent experiments; Pip, piperacillin; oxa, oxacillin; ctx, cefotaxim.
b. U: *S. oralis* Uo5; R: *S. pneumoniae* R6; the first letter refers to the promoter region, the second to the coding sequence of *murE*.

2007) and transferred to *S. pneumoniae* R6 as described in *Experimental procedures*. Two transformants each were used, and beta-galactosidase activity was assayed during late exponential growth and 2 h after reaching stationary phase. Each of the constructs with the *murE*$_{Uo5}$ promoter showed an approximately twofold higher activity with 80–98 Miller Units (MU) compared with 38–43 MU in case of the *murE*$_{R6}$ promoter at both time points.

To finally prove that the promoter of *murE*$_{Uo5}$ as well as its coding region can mediate an altered susceptibility to β-lactams, four derivatives of O*2x were constructed that contained an ectopic copy of the essential *murE* with different combinations of promoter versus coding sequence of *murE*$_{Uo5}$ respectively *murE*$_{R6}$. Subsequently the genomic *murE*$_{R6}$ was replaced with the spectinomycin resistance cassette *aad9* (LeBlanc *et al.*, 1991). The resulting strains were named according to their promoter and coding sequences: U for Uo5, R for R6, i.e. UO contains the promoter of *murE*$_{Uo5}$ and the coding regions of *murE*$_{R6}$ etc. All constructs containing *S. oralis* Uo5 sequences, UU, UR and RU, showed higher MICs compared with the RR strain with values similar to the transformants P20 and O(*2x*/*murE*) described above, especially for oxacillin and piperacillin (Table 1).

Taken together, these data prove that the promoter as well as the structural gene of *murE*$_{Uo5}$ contribute to beta-lactam resistance independently; the presence of both does not further increase resistance in the *S. pneumoniae* R6 genetic background.

MurE$_{Uo5}$ *contributes to resistance in combination with PBP2b*$_{Uo5}$ *and/or PBP2x*$_{Uo5}$

The capacity of MurE$_{Uo5}$ to confer resistance was also tested in combination with PBP2b$_{Uo5}$ and/or PBP2x$_{Uo5}$ using strains that were constructed exclusively with PCR fragments of the respective genes (Fig. 7). In each case, piperacillin was chosen to select for *pbp2b*$_{Uo5}$ and *murE*$_{Uo5}$. The presence of *murE*$_{Uo5}$ resulted primarily in a significant increase of piperacillin resistance compared with strains harboring the PBP genes alone (Table 2).

MurE$_{Uo5}$ *sequences in other resistant* Streptococcus *spp.*

If *murE*$_{Uo5}$ can contribute to beta-lactam resistance, one can expect to find relevant sequences, i.e. the promoter and internal gene regions, in other resistant streptococcal strains. Promoter regions almost identical to *murE*$_{Uo5}$ were found in some *S. oralis* unfinished genomes (*S. oralis* ATCC 49296 is shown as an example in Fig. 6) and *S. mitis* B6 (Fig. 6). Codons 113–183 are within the shortest fragment that was transferred from *S. oralis* Uo5 to *S. pneumoniae* R6; this region encodes the ATP binding motif. Such sequences were found also in *S. pneumoniae* genomes including strain *S. pneumoniae*

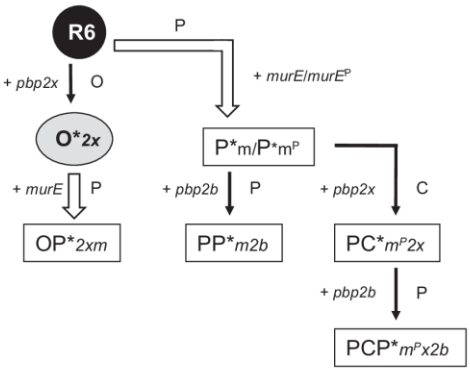


Fig. 7. *S. pneumoniae* R6 transformants containing *murE*$_{Uo5}$. Schematic representation of the selection steps. O (oxacillin), P (piperacillin) and C (cefotaxime) indicate the beta-lactams used for selection. Transformants are named according to the order of selective beta-lactams used during up to three transformation steps. The PCR amplified gene used as donor DNA is indicated by (+). *murE*^{P}: PCR fragment covering the promoter sequence. Genes of the resistance determinants that are present in the transformants are indicated by m (*murE*), 2x (*pbp2x*) and 2b (*pbp2b*). Transformation steps involving *murE*$_{Uo5}$ are indicated by open arrows.

Table 2. Beta-lactam susceptibilities of *S. pneumoniae* transformants with $murM_{Uo5}$ and/or PBP genes of *S. oralis* Uo5.

Strain	Recipient	Genotype	Selection (μg ml^{-1})	MIC (μg ml^{-1})[a] Pip	Oxa	Ctx
R6	–	WT	–	0.01	0.07	0.01
P*m	R6	$murE^{p}$/murE[b]	0.02 (Pip)	0.04–0.06	0.08–0.09	0.02–0.03
P*2b	R6	*pbp2b*	0.02 (Pip)	0.05–0.06	0.06–0.07 (ns)	0.02–0.03
PP*m2b	P*m	*murEpbp2b*	0.07 (Pip)	**0.2**	0.1–0.2	0.03
O*2x	R6	*pbp2x*	0.08 (Oxa)	0.01–0.02 (ns)	0.4–0.5	**0.2–0.3**
OP*2xm	O*	$pbp2xmurE^{p}$	0.02–0.03 (Pip)	**0.07–0.1**	**1.5**	**0.4–0.5**
OP*2x2b	O*	*pbp2xpbp2b*	0.03 (Pip)	**1–1.5**	**3–3.5**	**0.2–0.3**
PCP*m2x2b	PC*	$murE^{p}pbp2xpbp2b$	0.08–0.09 (Pip)	**3**	**3–4.5**	**0.4–0.5**

a. MIC values were taken from at least three independent experiments; Pip, piperacillin; oxa, oxacillin; ctx, cefotaxime. MICs which were at least 10-fold higher than that of the R6 strain are in bold letters and underlined; MICs that are less than twofold higher than that of R6 are marked with ns (not significant).
b. MIC values were identical for transformants containing the promoter region or coding sequences of $murE_{Uo5}$.
c. MIC values were identical for transformants containing nt 160–690, nt 300–1230, or the promoter regions plus the first 123 nt of $murE_{Uo5}$. MIC values that are significantly increased in individual transformants are in bold letters and underlined.

670, a representative of the clone Spain^{6B}-2 (Fig. 6), strongly suggesting that the MurE gene contributes to resistance also in these strains.

MurM in S. oralis *Uo5*

The $MurM_{Uo5}$ gene was not transferred in the three transformation steps with chromosomal *S. oralis* Uo5 DNA, i.e. was not present in the PCP genome. As *murM* represents a resistance determinant and has a mosaic structure in some penicillin clones of *S. pneumoniae* (Garcia-Bustos and Tomasz, 1990a; Smith and Klugman, 2001; Filipe *et al.*, 2002; Cafini *et al.*, 2005), we investigated $murM_{Uo5}$ in more detail.

The first curious finding was that the MurM gene is located in the *S. oralis* Uo5 genome at a different position compared with other *Streptococcus* spp. genomes. When compared with the *S. pneumoniae* R6 genome, the $murM_{Uo5}$ homologue *sor1963* is located in opposite orientation adjacent to degenerate transposase fragments corresponding to the position between *spr2020* and *spr2021*. This and the fact that $murM_{Uo5}$ differs by 41% in nt (corresponding to 50% aa) sequence from $murM_{R6}$ suggests that it has been imported into *S. oralis* Uo5 from an unknown source. Indeed, the closest homologue of $MurM_{Uo5}$ is MurM of *Streptococcus* sp. 140 (92.3% identity) according to BLAST analysis against the NCBI microbial genome data bank. A second interesting feature is the lack of a *murN* homologue in the *S. oralis* genome. These data suggest that the peptidoglycan of *S. oralis* has distinct interpeptide bridges in its peptidoglycan.

HPLC analysis combined with mass spectrometry of the *S. oralis* peptidoglycan indeed revealed several interesting features (Fig. 8, Table 3 and Table S2). First, branched muropeptides were abundant, and these contain one alanine attached to lysine instead of an Ala-Ala or Ser-Ala dipeptide found in *S. pneumoniae* (Garcia-Bustos *et al.*, 1987; Garcia-Bustos and Tomasz, 1990b; De Pascale *et al.*, 2008; Bui *et al.*, 2012), in agreement with the lack of *murN*. Although the MS analysis cannot distinguish whether the Ala resides at the epsilon amino group of Lys or at position 4 of the stem peptide, the elution time of the peaks in the HPLC analysis is consistent with branched muropeptides rather than with linear ones (Bui *et al.*, 2012). Second, some of the muropeptides are deacetylated in agreement with the presence of a homologue of the *pgdA* GlcNAc deacetylase gene (Vollmer and Tomasz, 2000) in the *S. oralis* Uo5 genome (*sor1343*). Third, some muropeptides contain an unamidated Glu residue instead of glutamine due to incomplete amidation mediated by the operon *gatD*/*murT*. This operon encodes the amidotransferase GatD and MurT with a substrate-binding domain of Mur ligases (Figueiredo *et al.*, 2012), homologues of which are present in *S. oralis* (*sor1448*/*sor1447*). A highly sensitive micro-LC-MS analysis revealed more than 50 distinct monomeric to tetrameric muropeptides and two further modifications resulting from the loss of sugar moieties (GlcNAc or GlcNAc-MurNAc) by the activities of glycosylhydrolases (Table S2).

Attempts to replace $murM_{R6}$ with $murM_{Uo5}$ in PCP via the Janus cassette (Sung *et al.*, 2001) proved to be difficult (not shown). Only one transformant could be obtained, and such a poor transformation efficiency strongly suggests that other compensatory mutations have occurred. $murM_{Uo5}$ could not be introduced in the absence of $murN_{R6}$ (not shown). This indicates that the presence of $murM_{Uo5}$ is not tolerated in the R6 background.

Discussion

The results presented here demonstrate that high beta-lactam resistance can be transferred from *S. oralis* into

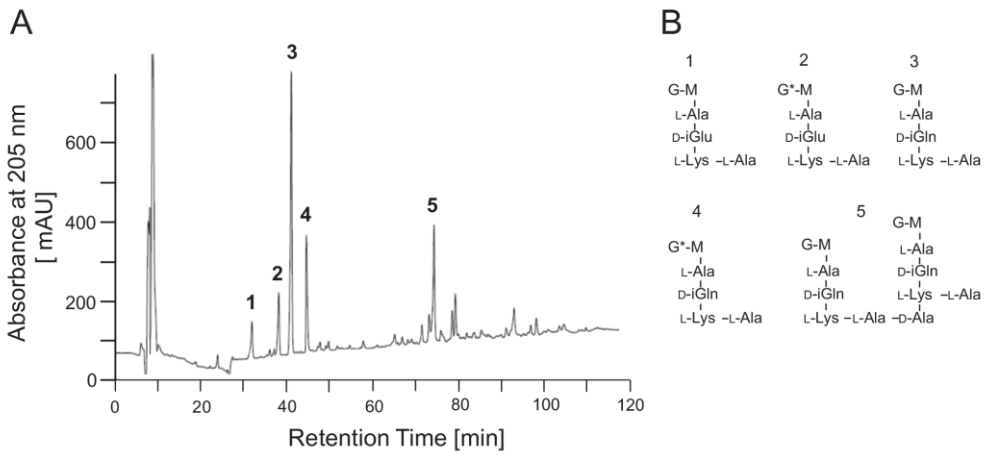


Fig. 8. Muropeptide profile of *S. oralis* Uo5 obtained by reversed-phase HPLC and proposed structures of major muropeptides.
A. Chromatogram showing reduced muropeptides from Uo5. Fractions number 1 to 5 were analysed by ESI-MS/MS (Table 3).
B. Proposed structures of muropeptides 1–5 from *S. oralis* Uo5. G, N-acetylglucosamine; G*, glucosamine, M, reduced N-acetylmuramic acid. Table S2 contains all muropeptides from Uo5 identified by micro-LC-MS analysis.

S. pneumoniae, involving four components important for peptidoglycan biosynthesis: the well-known primary penicillin resistance determinants and transpeptidases PBPs 2x and 2b; the bifunctional glycosyltransferase transpeptidase PBP1a; and surprisingly MurE, responsible for the addition of the lysine residue to peptidoglycan precursors. This combination adds a new aspect to the many pathways of resistance development in PSRP and related streptococci. Three of these components when present as the only *S. oralis* Uo5 gene in the *S. pneumoniae* R6 transformants confer resistance, albeit of low level and predominantly to different beta-lactams: PBP2x, PBP2b and MurE.

Table 3. Quantification and LTQ-FT MS analysis of the main muropeptides separated by HPLC.

Muropeptide[a]		Relative peak area (%)	Mass (amu) from collected peaks [H⁺]-form	
No.	Proposed structure		Theoretical	Determined
1	Tri(A)[Glu]	4.88	898.4257	898.4349
2	Tri(A) [Glu,deAc]	5.39	856.4151	856.4244
3	Tri(A)	24.05	897.4417	897.4511
4	Tri(A)[deAc]	8.91	855.4311	855.4458
5	Tri(A)Tetra(A)[Glu]	12.48	1847.8469	1847.9774

a. Tri(A), GlcNAc-MurNAc(r)-L-Ala-γ-D-Gln-L-Lys-ε-L-Ala; Tetra(A), GlcNAc-MurNAc(r)-L-Ala-γ-D-Gln-L-Lys(-ε-L-Ala)-D-Ala, with GlcNAc, N-acetylglucosamine; MurNAc(r), N-acetylmuramitol. [Glu] indicates an unamidated Glu residue instead of Gln; [deAc] indicates deacetylation of GlcNAc to glucosamine. Muropeptide No 5 is a dimer.

All three PBP genes are unique to *S. oralis* Uo5 and show signs of gene transfer when compared with those of the penicillin sensitive *S. oralis* ATCC35037 (Almaguer-Flores *et al.*, 2006) and other *Streptococcus* spp. genomes (Figs S1–S4), i.e. have a mosaic structure with sequences that occur in other *Streptococcus* spp. as well. $MurE_{Uo5}$ is highly similar to MurE in several *S. oralis* strains according to BLAST searches in the NCBI microbial draft genome database; however, it differs between codons 294 and 357 by almost 10% in DNA sequence, suggesting that it is a mosaic gene. Thus, all four resistance determinants appear to be acquired by *S. oralis* Uo5, which is known to develop competence at least under laboratory conditions (Reichmann *et al.*, 2011).

The total amount of DNA imported into *S. pneumoniae* R6 during three transformation steps (nine recombination events affecting 18.6 kb) was approximately 10-fold lower compared with *S. mitis* DNA acquired during four successive transformations (Sauerbier *et al.*, 2012). In addition to the regions that included the resistance determinants (altogether 10.7 kb), another five loci showed signs of recombination (Table S1). Their putative gene products did not suggest any role in resistance, but it remains to be clarified whether any of these genes affect the phenotype of the R6 transformants.

Alterations in PBPs

All R6 transformants expressing low affinity PBPs contained sequences that encode at least the transpeptidase domain of the respective PBPs, with mutations implicated in penicillin resistance (Figs S1–S3). PBP2x$_{Uo5}$ carries the mutation $T_{338}G$ (G_{339} in Uo5) next to the active site S_{337}, whereas it is A_{338} in most other related versions of PBP2x in PSRP. G_{338} confers a higher resistance level compared with A_{338} (Zerfaß *et al.*, 2009) and might thus contribute to the unusually high beta-lactam resistance of *S. oralis* Uo5. The effect of the other three sites that differ between Spain23F^{-1} and Uo5 ($I_{336}M$, $E_{378}D$ and $G_{384}S$) has not been investigated. PBP1a$_{Uo5}$, the other PBP implicated in cefotaxime resistance, contains the mutation $T_{371}S$ (S_{372} in Uo5) next to the active site S_{370}; an A_{371} occurs frequently in PSRP (Smith and Klugman, 1998; Asahi *et al.*, 1999; Ferroni and Berche, 2001; Nagai *et al.*, 2002; Nichol *et al.*, 2002) and is related to the resistance phenotype (Smith and Klugman, 1998; Nagai *et al.*, 2002; Job *et al.*, 2008). There are several other mutations that have been described in PBP1a of PSRP: $L_{583}M$, $A_{585}V$ and $P_{432}T$ (Smith and Klugman, 1998; Nichol *et al.*, 2002). The block of four altered amino acids $N_{574}TGY$ has been implicated in cephalosporin resistance (Smith and Klugman, 2003). This is present also in *S. oralis* ATCC35037, which is not a high level resistant strain (Almaguer-Flores *et al.*, 2006). It is conceivable that these alterations contribute to resistance only in the context of other mutations in PBP1a.

High level cefotaxime resistance occasionally required additional alterations in PBP2a (for review, see Hakenbeck *et al.*, 2012), but neither did we observe transfer of *pbp2a*$_{Uo5}$ nor were any significant changes such as $T_{441}A$ flanking the active site S_{440} apparent in PBP2a$_{Uo5}$ as described in PBP2a of some PSRP (Smith *et al.*, 2005). In contrast, the *S. mitis* B6 PBP2a was readily transferred during transformation experiments (Sauerbier *et al.*, 2012).

High level resistance to penicillin is mediated by a low affinity PBP2b$_{Uo5}$ in the transformants. Transfer of PBP2b into *S. pneumoniae* R6 appeared to be more difficult than that of the other PBP genes: transformation efficiency was always much lower, frequently only one mutation was transferred ($T_{446}A = A_{447}$ in Uo5 next to the $S_{443}SN$ motif), and regions downstream of nt 1350 corresponding to aa 450 of the protein were not detected in the various transformants analyzed (Fig. S3). This suggests that the $T_{446}A$ mutation is of major importance, whereas changes in the C-terminal region relate to *S. oralis* Uo5 specific features somehow involved in PBP2b$_{Uo5}$ structure or function. Failure to transform the entire PBP2b gene might be one of the reasons why the resistance levels of *S. oralis* Uo5 could not be reached in the transformants. Moreover, PBP2b$_{Uo5}$ contains one additional amino acid, D430, which is lacking in all other PSRP PBP2b variants. It is positioned on a loop between two short beta sheets β2b and β2c (Contreras-Martel *et al.*, 2009), adding a negative charge to the surface close to the active site cavity. Its impact on enzymatic activity cannot be deduced, but it could have an effect on the interaction with the substrate or partner molecules of PBP2b.

Low affinity PBP2x of PSRP has been shown to have a lower transpeptidase activity than those of their sensitive counterpart (Zhao *et al.*, 1997). It is likely that the mutations in the other low affinity PBPs, including those described above, also affect their activity that might have to be compensated by other components of the enzymatic machinery involved in peptidoglycan biosynthesis. It is possible that the special MurM of *S. oralis* Uo5 is important in this context. Most importantly, PBP2b function has been shown to somehow depend on the availability of branched muropeptides, the product of MurM (Berg *et al.*, 2013). Berg *et al.* showed that MurM-deficient mutants require much higher levels of PBP2b for growth compared with those that contain MurM (Berg *et al.*, 2013). The presence of MurM only, i.e. the absence of MurN seems to be common among *S. oralis* (own unpubl. data), and this might contribute to the difficulty to transfer beta-lactam resistance from *S. oralis* into *S. pneumoniae*.

MurE$_{Uo5}$ – alterations and resistance

The surprising result was not only that *murE*$_{Uo5}$ mediates resistance in the R6 background but that also the *murE*$_{Uo5}$ promoter region without any change in the coding sequence confers this phenotype. The *murE*$_{Uo5}$ promoter drives a twofold increase in expression, and this suggests that the structural alterations may lead to a higher activity of the enzyme. The region nt 300–687 represents the intragenic fragment that overlapped in all transformants analysed here, corresponding to alterations between aa 113 and 183, which include the ATP-binding motif G_{117}TKGK. The structure of *S. aureus* MurE in the presence of ADP and its reaction product UDP-MurNAc-L-Ala-γ-D-Glu-L-Lys has been solved (Ruane *et al.*, 2013). According to this structure, there are two sites that may affect the activity of MurE$_{Uo5}$: A126 (T126 in R6), which is located next to the ATP binding site, and the side-chains of S_{158}FS (A_{158}LT in R6) are directed toward the UDP-MurNAc-L-Ala-γ-D-Glu-L-Lys moiety. Unfortunately, attempts to overexpress MurE$_{Uo5}$ in *Escherichia coli* have not been successful preventing us to purify the enzyme to test its biochemical properties.

MurE has also been identified as an important factor for oxacillin resistance in *S. aureus* (Gardete *et al.*, 2004). In the MRSA strain used by Gardete *et al.*, an insertion 3 nt upstream the stop codon resulted in reduced specific activity of MurE and a drastic reduction especially of oxa-

cillin resistance in the mutant strain. Higher *murE* expression resulted in higher oxacillin resistance, similar to what we observed in the present study. It was also noted that a decline in *murE* transcription parallels decreased transcription of *pbpB* and *mecA* encoding PBP2 and the MRSA specific PBP2A respectively. *S. aureus* PBP2 is the homologue of *S. pneumoniae* PBP2x. We did not obtain any evidence that PBP2x is produced in higher amounts in the *S. pneumoniae* Δ*murE* derivatives containing either the promoter or coding region of $murE_{Uo5}$ using Western blots and anti-PBP2x antiserum, consistent with the unaltered expression of a *pbp2x* promotor-*lacZ* fusion (data not shown). In other words, within the twofold difference in *murE* expression that affects beta-lactam susceptibility in *S. pneumoniae*, the amount of PBP2x was not affected. It is possible that the phenomenon described in *S. aureus* is related to the different number of PBPs in this organism that contains only three hmw PBPs compared with four in *S. pneumoniae* and/or the presence of the additional PBP2a, which is specifically associated with MRSA strains. In this context the unusual cell wall biochemistry of *S. oralis* Uo5 is remarkable, with branched muropeptides containing only one Ala on the epsilon-amino groups of the stem peptide lysine residues, the consequence of a distinct MurM allele and the lack of MurN. It is conceivable that the combination of altered MurE and MurM in *S. oralis* ensures an optimal concentration of muropeptide precursors, the substrates for PBP-mediated transpeptidation reactions, for proper cell growth and division.

BLAST search of the NCBI microbial genome database revealed the presence of the $murE_{Uo5}$ promoter sequence highly similar to that of $MurE_{Uo5}$ in some *S. oralis* genomes, and in *S. mitis* B6 as well but not in *S. pneumoniae* genomes (Fig. 6). The alterations within MurE that are present in the $murE_{Uo5}$ transformants described here were also found in these genomes, as well as in five *S. pneumoniae* strains of different serotypes TCH8431/19A, A026, Taiwan19F-14, ST556 and 670-6B (Fig. 6). This clearly documents dissemination of the altered MurE gene among oral streptococci, but its origin remains unknown. $MurE_{Uo5}$ alone mediates only low level of resistance to beta-lactams but has a remarkable effect on resistance when present together with $PBP2x_{Uo5}$ or $PBP2b_{Uo5}$ (Table 2), thus facilitating the selection of such PBP variants during the evolution of beta-lactam resistance.

Experimental procedures

Bacterial strains and growth conditions

Streptococcus pneumoniae R6 is a beta-lactam-sensitive, unencapsulated laboratory strain derived from Rockefeller University strain R36A (Avery *et al.*, 1944). *S. oralis* Uo5 is a high level beta-lactam resistant isolate from Hungary from a nasal swab (Reichmann *et al.*, 1997). Streptococci were grown in C-medium (Lacks and Hotchkiss, 1960) supplemented with 0.2% yeast extract at 37°C without aeration or on blood agar plates (D-Agar; Alloing *et al.*, 1996) supplemented with 3% defibrinated sheep blood. Growth in liquid culture was monitored by nephelometry (nephelo units [NU]).

MIC determination

To determine minimal inhibitory concentrations (MICs), cultures of *S. pneumoniae* were grown in C-medium to a density of NU = 30, and after 1000-fold dilution in 0.9% NaCl aliquots (30 µl) were spotted on D-agar plates containing the antibiotic. Narrow concentrations at and above the MIC value of the recipient were used. MIC values were monitored after 24 incubation at 37°C. Strains were tested in at least three independent experiments. Antibiotic resistance genes used for chromosomal integrations in *S. pneumoniae* were selected with 2 µg ml^{-1} erythromycin (trimethoprime, 15 µg ml^{-1}; spectinomycin, 80 µg ml^{-1}).

Transformation

Transformation of *S. pneumoniae* was performed using competent cells as described previously (Mascher *et al.*, 2003). Transformation efficiency was calculated as the percentage of colonies obtained on the selective medium compared with the colony number on control plates without antibiotic. Chromosomal DNA from *S. pneumoniae* strains and *S. oralis* Uo5 was prepared as described previously (Laible *et al.*, 1989) except that *S. oralis* Uo5 was lysed in the presence of lysozyme (25 mg ml^{-1}) plus mutanolysin (0.5 mg ml^{-1}). PCR fragments were used as donor DNA after purification from agarose gels. Transformants were named according to the order of selective beta-lactams (O, oxacillin; P, piperacillin; C, cefotaxime). Transformants obtained with PCR fragments are indicated by asterisk (*). At least 10 transformants from each experiment were analysed for MICs and PBP profiles, and at least two transformants were selected for DNA sequencing of the gene encoding a low affinity PBP. In case of $murE_{Uo5}$ transformants, *murE* was sequenced in at least three of those displaying a higher MIC than the recipient. One transformant was chosen as recipient for further transformation experiments.

Amplification and sequencing of PBP genes

Amplification of chromosomal DNA with PCR was carried out using either GoldStar *Taq* polymerase or high fidelity iProof polymerase according to the manufacturer's instructions. Alternatively, 100 µl cells from glycerol stock were centrifuged and resuspended in 50 µl H_2O; 1 µl of this suspension was used directly in a PCR reaction. DNA fragments were purified using an extraction kit (NucleoSpin®Extract II; Macherey-Nagel, Düren) as described by the manufacturer. The presence of PBP and MurE sequences in the transformants was confirmed by DNA sequencing. The oligonucleotides used for PCR- and sequencing-reactions relevant for this study are listed in Table S3.

Detection of penicillin-binding proteins

Cells of an exponentially growing culture were harvested by centrifugation and resuspended in 20 mM sodium phosphate

buffer pH 7.2. The volume was adjusted so that 5 µl cell suspension corresponds to 1 ml culture at NU = 20. Cells were lysed in 20 mM sodium phosphate buffer pH 7.2 containing 0.2% Triton X-100 during an incubation of 30 min at 37°C. *S. oralis* Uo5 cells were lysed after addition of lysozyme (0.8 mg ml^{-1}) and cellosyl (0.5 mg ml^{-1}) in the same buffer. For labelling, 5 µl of the cell lysate were incubated with 3 µl of Bocillin™FL (10 µM) for approximately 20 min (Zhao *et al.*, 1999). Proteins were separated by SDS-polyacrylamide gel electrophoresis (PAGE). The final acrylamide concentration of the separation gel was 7.5%, the ratio of acrylamide : bisacrylamide = 30:0.8. Bocillin–PBP complexes were visualised with a FluorImager at 488 nm (GE-Healthcare). Proteins were stained with Coomassie brilliant blue.

Construction of an ectopic copy of murE

Integration of a copy of *murE* in the *S. pneumoniae* transformant O*2x was achieved by means of the integrative plasmid pSW1, which can be selected with trimethoprime (Denapaite and Hakenbeck, 2011). The genes *bgaA* and *spr0566-spr0568* serve as recombination sites for integration into the *S. pneumoniae* genome. PCR fragments from chromosomal DNA were generated using the primer pair PM 278/PM270 and the PM 277/PM270 for the amplification of the *S. oralis* Uo5 respectively *S. pneumoniae* R6 promotor region of *murE*. After digestion with *Bam*HI and *Nhe*I, the amplified fragments were ligated with *Bam*HI-*Nhe*I-digested pSW1 vector DNA. The ligation mixtures were transformed into strain O*2x, and transformants were selected with 15 µg ml^{-1} trimethoprim. For the inactivation of wild-type *murE* in the strains constructed above, the spectinomycin-resistance gene *aad9* was amplified from chromosomal DNA of *S. pneumoniae* strain *ciaR*::*aad9* (Mascher *et al.*, 2003) using the primer pair PM273/PM274. The fragment was digested with BamHI and NheI. Flanking regions of approximately 1 kb down- and upstream of $murE_{R6}$ were amplified with the primer pairs PM272/PM171 and PM170/PM271 and digested with NheI and BamHI respectively. After ligation and transformation of the ligation mixture, transformants with inactivated *murE* were selected with 80 µg ml^{-1} spectinomycin.

Construction of reporter plasmids

For cloning the *murE* promoter fragments into pPP2 (Halfmann *et al.*, 2007), the promoter regions from *murE* of *S. pneumoniae* R6 and *S. oralis* Uo5 were amplified with the PCR primer pair PM275/PM276. The PCR products were cleaved with SphI and BamHI and ligated into pPP2 digested with the same enzymes. The plasmids were transferred to *S. pneumoniae* R6 using the tetracycline resistance marker *tetM* for selection as described previously (Halfmann *et al.*, 2007).

Determination of β-galactosidase activity

Cell extracts were prepared from of *S. pneumoniae* cultures grown to a density of NU = 80; cell lysis and determination of specific β-galactosidase activities were performed as described (Halfmann *et al.*, 2007). β-galactosidase assays were performed in six independent experiments.

DNA microarray analysis

The oligonucleotide microarray covering genes and intergenic regions of *S. pneumoniae* R6/TIGR4 is listed under ArrayExpress accession number A-MEXP-1846. Oligonucleotides were spotted on Nexterion HiSens Slides E (SCHOTT Jenaer Glas GmbH) using the SpotArray TM24 Microarray Spotting System (Perkin-Elmer) with 32 SMP3-Pins (Telechem). DNA labeling, hybridization and data processing were performed as described (Sauerbier *et al.*, 2012).

Cell wall analysis

Preparation of *S. pneumoniae* R6 and *S. oralis* Uo5 cell walls and peptidoglycan, digestion of peptidoglycan with cellosyl, separation of muropeptides by HPLC and MS analysis were performed essentially as described recently (Bui *et al.*, 2012). In brief, cells of a late exponential phase culture were collected by centrifugation. After boiling in 5% sodium dodecyl sulfate and washing, cells were disrupted with glass beads. Cell walls were treated with 48% hydrofluoric acid to remove secondary cell wall polymers, and the resulting peptidoglycan was digested with cellosyl for 18 h. After reduction with $NaBH_4$, muropeptides were separated by HPLC on a reversed-phase column. Muropeptides were detected at 205 nm, and collected fractions of major muropeptides were analysed at the Newcastle University Pinnacle facility by offline electrospray-MS. For details, see Bui *et al.* (2012). In addition, non-reduced muropeptides were analysed by online micro-HPLC–MS as described (Bui *et al.*, 2009).

DNA sequencing

The genome of the transformant PCP was sequenced using illumina sequencing technology and paired end information, and reads were assembled with paired end information (~ 600 bp distance) by the Newbler gsAssembler Version 2.6. Contigs were aligned to the *S. pneumoniae* R6 sequence (Lanie *et al.*, 2007) and to *S. oralis* Uo5 (Reichmann *et al.*, 2011), and regions of identity to *S. oralis* Uo5 were identified using BLAST analyses in combination with visual inspection of the regions.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft Ha1011/11-1, Ha1011/11-2 (to R.H.), and the EC through the DIVINOCELL project (to W.V.). We thank Ulrike Klein and Cathrin Schmeiser for help in MIC determination and DNA sequencing, Katja Frohnweiler for help in the construction of *murM* derivatives and Reinhold Brückner for helpful discussions. The authors have no conflict of interest to declare.

References

Alloing, G., Granadel, C., Morrison, D.A., and Claverys, J.-P. (1996) Competence pheromone, oligopeptide permease,

and induction of competence in *Streptococcus pneumoniae*. *Mol Microbiol* **21:** 471–478.
Almaguer-Flores, A., Moreno-Borjas, J.Y., Salgado-Martinez, A., Sanchez-Reyes, M.A., Alcantara-Maruri, E., and Ximenez-Fyvie, L.A. (2006) Proportion of antibiotic resistance in subgingival plaque samples from Mexican subjects. *J Clin Periodontol* **33:** 743–748.
Asahi, Y., Takeuchi, Y., and Ubukata, K. (1999) Diversity of substitutions within or adjacent to conserved amino acid motifs of penicillin-binding protein 2x in cephalosporin-resistant *Streptococcus pneumoniae* isolates. *Antimicrob Agents Chemother* **43:** 1252–1255.
Avery, O.T., MacLeod, C.M., and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med* **79:** 137–158.
Berg, K.H., Stamsas, G.A., Straume, D., and Havarstein, L.S. (2013) The effect of low Pbp2b levels on cell morphology and peptidoglycan composition in *Streptococcus pneumoniae* R6. *J Bacteriol* **195:** 4342–4354.
Bui, N.K., Gray, J., Schwarz, H., Schumann, P., Blanot, D., and Vollmer, W. (2009) The peptidoglycan sacculus of *Myxococcus xanthus* has unusual structural features and is degraded during glycerol-induced myxospore development. *J Bacteriol* **191:** 494–505.
Bui, N.K., Eberhardt, A., Vollmer, D., Kern, T., Bougault, C., Tomasz, A., *et al.* (2012) Isolation and analysis of cell wall components from *Streptococcus pneumoniae*. *Anal Biochem* **421:** 657–666.
Cafini, F., del Campo, R., Alou, L., Sevillano, D., Morosini, M.I., Baquero, F., *et al.* (2005) Alterations of the penicillin-binding proteins and *murM* alleles of clinical *Streptococcus pneumoniae* isolates with high-level resistance to amoxicillin in Spain. *J Antimicrob Chemother* **57:** 224–229.
del Campo, R., Cafini, F., Morosini, M.I., Fenoll, A., Liñares, J., Alou, L., *et al.* (2006) Combinations of PBPs and MurM protein variants in early and contemporary high-level penicillin-resistant *Streptococcus pneumoniae* isolates in Spain. *J Antimicrob Chemother* **57:** 983–986.
Contreras-Martel, C., Dahout-Gonzalez, C., Martins Ados, S., Kotnik, M., and Dessen, A. (2009) PBP active site flexibility as the key mechanism for beta-lactam resistance in pneumococci. *J Mol Biol* **387:** 899–909.
De Pascale, G., Lloyd, A.J., Schouten, J.A., Gilbey, A.M., Roper, D.I., Dowson, C.G., and Bugg, T.D. (2008) Kinetic characterization of lipid II-Ala:alanyl-tRNA ligase (MurN) from *Streptococcus pneumoniae* using semisynthetic aminoacyl-lipid II substrates. *J Biol Chem* **283:** 34571–34579.
Denapaite, D., and Hakenbeck, R. (2011) A new variant of the capsule 3 cluster occurs in *Streptococcus pneumoniae* from deceased wild chimpanzees. *PLoS ONE* **6:** e25119.
Dowson, C.G., Hutchison, A., and Spratt, B.G. (1989) Extensive re-modelling of the transpeptidase domain of penicillin-binding protein 2B of a penicillin-resistant South African isolate of *Streptococcus pneumoniae*. *Mol Microbiol* **3:** 95–102.
Dowson, C.G., Coffey, T.J., Kell, C., and Whiley, R.A. (1993) Evolution of penicillin resistance in *Streptococcus pneumoniae*; the role of *Streptococcus mitis* in the formation of a low affinity PBP2B in *S. pneumoniae*. *Mol Microbiol* **9:** 635–643.
Ferroni, A., and Berche, P. (2001) Alterations to penicillin-binding proteins 1A, 2B and 2X amongst penicillin-resistant clinical isolates of *Streptococcus pneumoniae* serotype 23F from the nasopharyngeal flora of children. *J Med Microbiol* **50:** 828–832.
Figueiredo, T.A., Sobral, R.G., Ludovice, A.M., Almeida, J.M., Bui, N.K., Vollmer, W., *et al.* (2012) Identification of genetic determinants and enzymes involved with the amidation of glutamic acid residues in the peptidoglycan of *Staphylococcus aureus*. *PLoS Pathog* **8:** e1002508.
Filipe, S., Severina, E., and Tomasz, A. (2001) Distribution of the mosaic structured *murM* genes among natural populations of *Streptococcus pneumoniae*. *J Bacteriol* **182:** 6798–6805.
Filipe, S.R., and Tomasz, A. (2000) Inhibition of the expression of penicillin-resistance in *Streptococcus pneumoniae* by inactivation of cell wall muropeptide branching genes. *Proc Natl Acad Sci USA* **97:** 4891–4896.
Filipe, S.R., Severina, E., and Tomasz, A. (2002) The *murMN* operon: a functional link between antibiotic resistance and antibiotic tolerance in *Streptococcus pneumoniae*. *Proc Natl Acad Sci USA* **99:** 1550–1555.
Garcia-Bustos, J., and Tomasz, A. (1990a) A biological price of antibiotic resistance: major changes in the peptidoglycan structure of penicillin-resistant pneumococci. *Proc Natl Acad Sci USA* **87:** 5415–5419.
Garcia-Bustos, J., and Tomasz, A. (1990b) A biological price of antibiotic resistance: major changes in the peptidoglycan structure of penicillin-resistant pneumococci. *Proc Natl Acad Sci USA* **87:** 5415–5419.
Garcia-Bustos, J.F., Chait, B.T., and Tomasz, A. (1987) Structure of the peptide network of pneumococcal peptidoglycan. *J Biol Chem* **262:** 15400–15405.
Gardete, S., Ludovice, A.M., Sobral, R.G., Filipe, S.R., de Lencastre, H., and Tomasz, A. (2004) Role of *murE* in the expression of beta-lactam antibiotic resistance in *Staphylococcus aureus*. *J Bacteriol* **186:** 1705–1713.
Goffin, C., and Ghuysen, J.-M. (1998) Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. *Microbiol Mol Biol Rev* **62:** 1079–1093.
Grebe, T., and Hakenbeck, R. (1996) Penicillin-binding proteins 2b and 2x of *Streptococcus pneumoniae* are primary resistance determinants for different classes of β-lactam antibiotics. *Antimicrob Agents Chemother* **40:** 829–834.
Hakenbeck, R., Tornette, S., and Adkinson, N.F. (1987) Interaction of non-lytic β-lactams with penicillin-binding proteins in *Streptococcus pneumoniae*. *J Gen Microbiol* **133:** 755–760.
Hakenbeck, R., König, A., Kern, I., van der Linden, M., Keck, W., Billot-Klein, D., *et al.* (1998) Acquisition of five high-M, penicillin-binding protein variants during transfer of high-level β-lactam resistance from *Streptococcus mitis* to *Streptococcus pneumoniae*. *J Bacteriol* **180:** 1831–1840.
Hakenbeck, R., Brückner, R., Denapaite, D., and Maurer, P. (2012) Molecular mechanism of beta-lactam resistance in *Streptococcus pneumoniae*. *Future Microbiol* **7:** 395–410.
Halfmann, A., Hakenbeck, R., and Brückner, R. (2007) A new integrative reporter plasmid for *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **268:** 217–224.

Henriques-Normark, B. (2007) Molecular epidemiology and mechanisms for antibiotic resistance in *Streptococcus pneumoniae*. In *Molecular Biology of Streptococci*. Hakenbeck, R., and Chhatwal, G.S. (eds). Wymondham, Norfolk, UK: Horizon Press, pp. 269–290.

Job, V., Carapito, R., Vernet, T., Dessen, A., and Zapun, A. (2008) Common alterations in PBP1a from resistant *Streptococcus pneumoniae* decrease its reactivity toward beta-lactams: structural insights. *J Biol Chem* **283:** 4886–4894.

Lacks, S., and Hotchkiss, R.D. (1960) A study of the genetic material determining an enzyme activity in pneumococcus. *Biochim Biophys Acta* **39:** 508–517.

Laible, G., Hakenbeck, R., Sicard, M.A., Joris, B., and Ghuysen, J.-M. (1989) Nucleotide sequences of the *pbpX* genes encoding the penicillin-binding protein 2x from *Streptococcus pneumoniae* R6 and a cefotaxime-resistant mutant, C506. *Mol Microbiol* **3:** 1337–1348.

Laible, G., Spratt, B.G., and Hakenbeck, R. (1991) Interspecies recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Mol Microbiol* **5:** 1993–2002.

Lanie, J.A., Ng, W.L., Kazmierczak, K.M., Andrzejewski, T.M., Davidsen, T.M., Wayne, K.J., *et al.* (2007) Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. *J Bacteriol* **297:** 503–512.

LeBlanc, D.J., Lee, L.N., and Inamine, J.M. (1991) Cloning and nucleotide base sequence analysis of a spectinomycin adenyltransferase AAD(9) determinant from *Enterococcus faecalis*. *Antimicrob Agents Chemother* **35:** 1804–1810.

Lloyd, A.J., Gilbey, A.M., Blewett, A.M., De, P.G., El, Z.A., Levesque, R.C., *et al.* (2008) Characterization of tRNA-dependent peptide bond formation by MurM in the synthesis of *Streptococcus pneumoniae* peptidoglycan. *J Biol Chem* **283:** 6402–6417.

Martin, C., Sibold, C., and Hakenbeck, R. (1992) Relatedness of penicillin-binding protein 1a genes from different clones of penicillin-resistant *Streptococcus pneumoniae* isolated in South Africa and Spain. *EMBO J* **11:** 3831–3836.

Mascher, T., Merai, M., Balmelle, N., de Saizieu, A., and Hakenbeck, R. (2003) The *Streptococcus pneumoniae cia* regulon: CiaR target sites and transcription profile analysis. *J Bacteriol* **185:** 60–70.

Muñóz, R., Dowson, C.G., Daniels, M., Coffey, T.J., Martin, C., Hakenbeck, R., and Spratt, B.G. (1992) Genetics of resistance to third-generation cephalosporins in clinical isolates of *Streptococcus pneumoniae*. *Mol Microbiol* **6:** 2461–2465.

Müller, M., Marx, P., Hakenbeck, R., and Brückner, R. (2011) Effect of new alleles of the histidine kinase gene *ciaH* on the activity of the response regulator CiaR in *Streptococcus pneumoniae* R6. *Microbiology* **157:** 3104–3112.

Nagai, K., Davies, T.A., Jacobs, M.R., and Appelbaum, P.C. (2002) Effects of amino acid alterations in penicillin-binding proteins (PBPs) 1a, 2b, and 2x on PBP affinities of penicillin, ampicillin, amoxicillin, cefditoren, cefuroxime, cefprozil, and cefaclor in 18 clinical isolates of penicillin-susceptible, -intermediate, and -resistant pneumococci. *Antimicrob Agents Chemother* **46:** 1273–1280.

Nichol, K.A., Zhanel, G.G., and Hoban, D.J. (2002) Penicillin-binding protein 1A, 2B, and 2X alterations in Canadian isolates of penicillin-resistant *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **46:** 3261–3264.

Reichmann, P., König, A., Liñares, J., Alcaide, F., Tenover, F.C., McDougal, L., *et al.* (1997) A global gene pool for high-level cephalosporin resistance in commensal *Streptococcus* spp. and *Streptococcus pneumoniae*. *J Infect Dis* **176:** 1001–1012.

Reichmann, P., Nuhn, M., Denapaite, D., Brückner, R., Henrich, B., Maurer, P., *et al.* (2011) Genome of *Streptococcus oralis* strain Uo5. *J Bacteriol* **193:** 2888–2889.

Ruane, K.M., Lloyd, A.J., Fulop, V., Dowson, C.G., Barreteau, H., Boniface, A., *et al.* (2013) Specificity determinants for lysine incorporation in *Staphylococcus aureus* peptidoglycan as revealed by the structure of a MurE enzyme ternary complex. *J Biol Chem* **288:** 33439–33448.

Sauerbier, J., Maurer, P., Rieger, M., and Hakenbeck, R. (2012) *Streptococcus pneumonia*e R6 interspecies transformation: genetic analysis of penicillin resistance determinants and genome-wide recombination events. *Mol Microbiol* **86:** 692–706.

Sauvage, E., Kerff, F., Terrak, M., Ayala, J., and Charlier, P. (2008) The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS Microbiol Rev* **32:** 234–258.

Severin, A., and Tomasz, A. (1996) Naturally occurring peptidoglycan variants of *Streptococcus pneumoniae*. *J Bacteriol* **178:** 168–174.

Sibold, C., Henrichsen, J., König, A., Martin, C., Chalkley, L., and Hakenbeck, R. (1994) Mosaic *pbpX* genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from *pbpX* genes of a penicillin-sensitive *Streptococcus oralis*. *Mol Microbiol* **12:** 1013–1023.

Smith, A.M., and Klugman, K.P. (1998) Alterations in PBP1A essential for high-level penicillin resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **42:** 1329–1333.

Smith, A.M., and Klugman, K.P. (2001) Alterations in MurM, a cell wall muropeptide branching enzyme, increase high-level penicillin and cephalosporin resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **45:** 2393–2396.

Smith, A.M., and Klugman, K.P. (2003) Site-specific mutagenesis analysis of PBP 1A from a penicillin-cephalosporin-resistant pneumococcal isolate. *Antimicrob Agents Chemother* **48:** 387–389.

Smith, A.M., Feldman, C., Massidda, O., McCarthy, K., Ndiweni, D., and Klugman, K.P. (2005) Altered PBP 2A and its role in the development of penicillin, cefotaxime, and ceftriaxone resistance in a clinical isolate of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **49:** 2002–2007.

Sung, C.K., Li, H., Claverys, J.P., and Morrison, D.A. (2001) An *rpsL* cassette, janus, for gene replacement through negative selection in *Streptococcus pneumoniae*. *Appl Environ Microbiol* **67:** 5190–5196.

Vollmer, W., and Tomasz, A. (2000) The pgdA gene encodes for a peptidoglycan N-acetylglucosamine deacetylase in *Streptococcus pneumoniae*. *J Biol Chem* **275:** 20496–20501.

Weber, B., Ehlert, K., Diehl, A., Reichmann, P., Labischinski, H., and Hakenbeck, R. (2000) The *fib* locus in *Streptococcus pneumoniae* is required for peptidoglycan crosslinking and PBP-mediated beta-lactam resistance. *FEMS Microbiol Lett* **188:** 81–85.

Zerfaß, I., Hakenbeck, R., and Denapaite, D. (2009) An important site in PBP2x of penicillin-resistant clinical isolates of *Streptococcus pneumoniae*: mutational analysis of Thr338. *Antimicrob Agents Chemother* **53:** 1107–1115.

Zhao, G., Yeh, W.-K., Carnahan, R.H., Flokowitsch, J., Meier, T.I., Alborn, W.E., Jr, *et al.* (1997) Biochemical characterization of penicillin-resistant and -sensitive penicillin-binding protein 2x transpeptidase activities of *Streptococcus pneumoniae* and mechanistic implications in bacterial resistance to β-lactam antibiotics. *J Bacteriol* **179:** 4901–4908.

Zhao, G., Meir, T.I., Kahl, S.D., Gee, K.R., and Blaszczak, L.C. (1999) Bocillin FL, a sensitive and commercially available reagent for detection of penicillin-binding proteins. *Antimicrob Agents Chemother* **43:** 1124–1128.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

2.5 Genomics, genetic variation, and regions of differences

Hervé Tettelin, Scott Chancey, Tim Mitchell, Dalia Denapaite, Yvonne Schähle, **Martin Rieger** and Regine Hakenbeck. In: Jeremy Brown, Sven Hammerschmidt and Carlos Orihuela. ***Streptococcus pneumoniae*: Molecular mechanisms of host-pathogen interactions**. 2015 May; Elsevier Science Publishing Co Inc. ISBN: 978-0-12-410530-0

Summary:

This book chapter concerns the genomics of the bacterial species *Streptococcus pneumoniae*. The three parts address questions related to whole genome analysis, virulence factors, and differences to closely related commensal species.

Several studies revealed genomic alterations as a result of horizontal transfer, such as capsule switching and acquisition of antibiotic resistance in response to clinical interventions. These studies demonstrate the ability of whole genome sequencing and comparative analysis to generate deeper insights into the evolution of the species *S. pneumoniae*.

The pan-genome of *S. pneumoniae*, the repertoire of genes accessible to this species, is quite large. A new pan-genome analysis resulted in a new formula to predict new genes found within a given number of new genome sequences. The core-genome comprises only genes shared by all strains of the species and which are required for basic functions. However, depending on the method and the genomes used, the number of core genes varies between ~950 - 1.100 with a total number of genes around 2.100. In contrast, the dispensable genome (approximately 25% of the pneumococcal genome) which is present only in a subset of strains, comprises a huge diversity and can provide advantages such as antibiotic resistance and variable host defence mechanisms.

Multi locus sequence typing (MLST), based on comparison of housekeeping genes, is the current typing method of choice for the definition of clones. Comparative genomics has revealed a substantial variation within clones defined by MLST mainly due to horizontal gene transfer events. *S. pneumoniae* is well adapted to gain new DNA by transformation and

recombination. Alternative ways of gene transfer involve integrative and conjugative element (ICE), phages and insertion sequence (IS) which contribute to an increase of the pan-genome.

The role and variability of pneumococcal virulence factors such as the polysaccharide capsule, surface proteins and the cytolysin pneumolysin will be discussed in the second part. In addition, the role of two-component systems (TCS) that mediate physiological responses to environmental signals for virulence, and the impact of the variable genetic background on infection potential is discussed.

The third part addresses differences between the pathogen *S. pneumoniae* and its close relatives, commensal streptococci. Genomic comparison documents many examples of horizontal gene transfer between species, and examples of gene clusters that occur in several species are documented. The differentiation between the species can be achieved by comparative analyses of house-keeping genes (MLST and MLSA), but genomic hybridisation data revealed a smooth transition between the species, due to the large dispensable genome which is circulating among different species. Finished genomes are available for *S. pneumoniae* R6, *S. mitis* B6 and *S. oralis* Uo5, and core genome analysis revealed that about 60% of the deduced proteins are common among these strains. Of the 532 proteins, which are specific to R6 in this analysis, only 104 remain *S. pneumoniae*-specific after comparison with another 26 pneumococcal genomes. These genes probably include factors important for the adaption to the ecological niche and the pathogenicity potential of *S. pneumoniae*. The MLST tree of *S. pneumoniae*, *S. mitis*, *S. oralis* and *S. pseudopneumoniae* reveals a common ancestor of *S. pneumoniae* and *S. mitis* with later diversification of *S. pneumoniae*, probably in parallel to human evolution (Kilian, et al., 2008). Genomic analyses showed that horizontal gene transfer occurred mainly unidirectional from *S. mitis* to *S. pneumoniae*. This is supported by analysis of capsular genes and mosaic genes (e.g. PBPs) as well as the presence of disrupted versus complete genes in these two species.

Most virulence factors of *S. pneumoniae* including many surface proteins can be found in other Mitis-group streptococci. On the other hand, the two-component system TCS06 and the variable choline-binding proteins *pspA*, *pcpA* and *pspC* can be considered to be specific for *S. pneumoniae*. This is also true for the hyaluronidase *hlyA* and the pneumo-/autolysin-island (*ply-lytA*) which occur rarely in some strains of other species. The highly variable capsule cluster is essential for pneumococcal virulence but it was recently found, that virtually all

commensal viridans streptococci are capable of capsule expression (Kilian, et al., 2019; Skov Sørensen, et al., 2016). Genomic comparison of representatives of the species *S. pneumoniae*, *S. mitis* and *S. oralis* reveals a core genome of similar size (1.140 genes) compared to the core genome of *S. pneumoniae*. In summary, this chapter emphasizes the progress in our understanding of the pneumococcal biology in the genomic area.

Own contribution to the paper:

Estimation of core and accessory genomes of *S. pneumoniae* R6, *S. mitis* B6, *S. oralis* Uo5 and *S. pseudopneumoniae*. Estimation of *S. pneumoniae* R6-specific genes as well as core and accessory genome of 29 further *S. pneumoniae* genomes as described in chapter 3.5.

CHAPTER

5

Genomics, Genetic Variation, and Regions of Differences

Hervé Tettelin[1], *Scott Chancey*[2], *Tim Mitchell*[3], *Dalia Denapaite*[4], *Yvonne Schähle*[4], *Martin Rieger*[4] *and Regine Hakenbeck*[4]

[1]Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA [2]Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, and Laboratories of Microbial Pathogenesis, Department of Veterans Affairs Medical Center, Atlanta, GA, USA [3]School of Immunity and Infection, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, UK [4]Department of Microbiology, University of Kaiserslautern, Kaiserslautern, Germany

STREPTOCOCCUS PNEUMONIAE COMPARATIVE GENOMICS

High-throughput sequencing technologies applied to bacterial pathogens provide an information-rich environment for the study of bacterial evolution, epidemiology, and pathogenesis [1]. Since publication of its first whole genome sequences in 2001 [2,3], *Streptococcus pneumoniae* has received a great deal of attention from a genomics perspective, with more than 4000 genomes sequenced to date, most of them in draft form, and many more to come. Comparative genomics analyses range from the comparison of assembled and annotated genomes to the mapping of unannotated raw sequence data to a reference genome.

A recent analysis of 240 *S. pneumoniae* isolates of the PMEN1 (Spain^{23F}-1) multidrug-resistant lineage has provided the ability to distinguish accumulated base substitutions from polymorphisms generated by recombination of imported DNA [4]. A phylogeny constructed using only vertically inherited single nucleotide polymorphisms (SNPs) (regions affected by recombination events were removed) proved to be a superior estimation of the evolution of the PMEN1 clone based on the correlation between the distance from the root of the tree and the date of isolation of each strain. By comparing this more accurate phylogeny to the dates and locations of isolation, the authors were able to pinpoint the likely origins of different clades and to discern information about the geographic

Streptococcus pneumoniae.
DOI: http://dx.doi.org/10.1016/B978-0-12-410530-0.00005-3

spread of the clone [5]. The Croucher et al. [4] study of closely related isolates from 22 countries revealed capsule switching events and acquisition of antibiotic resistance determinants in response to clinical interventions over short timescales. For instance, it identified 10 cases where PMEN1 isolates escaped pressure applied by the PCV7 vaccine by switching to the expression of non-vaccine serotypes. Analysis of integrative and conjugative mobile elements revealed selection for determinants of resistance to drugs commonly used for treatment of upper respiratory tract infections.

In a separate study, the genomes of 426 isolates from a genetically diverse, historical, and global collection covering the years 1937–2007 were sequenced in order to track the rapid increase of penicillin resistance [6]. A particular focus on PMEN1 revealed that its likely ancestor is one of the earliest known penicillin-nonsusceptible strains, isolated in 1967 in Australia. The study further indicated that PMEN1 is a very efficient donor of penicillin and other antibiotic resistance genes to many genotypically diverse *S. pneumoniae* lineages. This led the authors to designate PMEN1 a "paradigm for genetic success."

In-depth genomic analysis of the PMEN1 clone provided a detailed overview of the worldwide distribution and evolution of a set of closely related isolates that successfully adapted to host and therapeutic pressures. An alternate way of studying pneumococcal evolution is to compare unrelated isolates recovered over time from a specific geographical location. We took this approach in order to characterize the genomic diversity of *S. pneumoniae* clinical isolates within the Atlanta, Georgia, metropolitan area. Our study focused on the generation of 147 whole genome sequences, including 121 invasive and 10 carriage isolates from Atlanta as well as 16 invasive isolates from outside of Atlanta. The genomes encompassed 22 serotypes, 86 multilocus sequence typing (MLST) types, resistance to 10 antibiotics, and 10 disease outcomes. The collection included 29 strains belonging to the MLST-based clonal complex CC320, a multidrug-resistant complex responsible for the global emergence of non-vaccine serotype 19A in the years following the introduction of PCV7. The predicted founder of the complex, ST320, was a serotype 19A clone carrying dual macrolide resistance determinants (Mega and *erm*(B)) [7]. The clone represented a 19F-19A capsule switch and horizontal acquisition of multiple antibiotic resistance mechanisms, suggesting that vaccine and antibiotic pressures influenced its emergence [8].

The 29 CC320 whole genome sequences in our study (20 Atlanta invasive, 1 Atlanta carriage, and 8 invasive isolates from states outside of Georgia), together with publicly available CC320 genomes TCH8431/19A (ST320) and Taiwan^{19F}-14 (ST236), were subjected to whole genome multiple sequence alignment using the Mugsy software [9] with default parameters. The alignment in MAF format was then filtered with Phylomark [10] to extract and concatenate the core nucleotides, including SNPs, and to construct a neighbor-joining phylogenetic tree using MEGA v6.06 [11]. Of the 31 CC320 isolates, 18 were serotype 19F and 13 were 19A. In order to eliminate the influence of capsule switching on the phylogeny, we deleted the capsule locus from the genome sequence of the Taiwan^{19F}-14 isolate such that the capsule locus would no longer be part of the core alignment. The genomes clustered into three clades representing each of the three subgroups in CC320: ST236, ST271, and ST320 (Figure 5.1). Clade 236 consisted exclusively of serotype 19F isolates belonging to the CC320 subgroup founded by ST236. Clade 236 isolates were identified in the pre- and post-vaccine eras, despite vaccine pressure against serotype 19F due to PCV7. This could be explained by antibiotic pressure selecting for the multidrug-resistant phenotype. Clade 236 isolates, with exception of one pre-PCV7 era isolate (GA13499), were resistant to

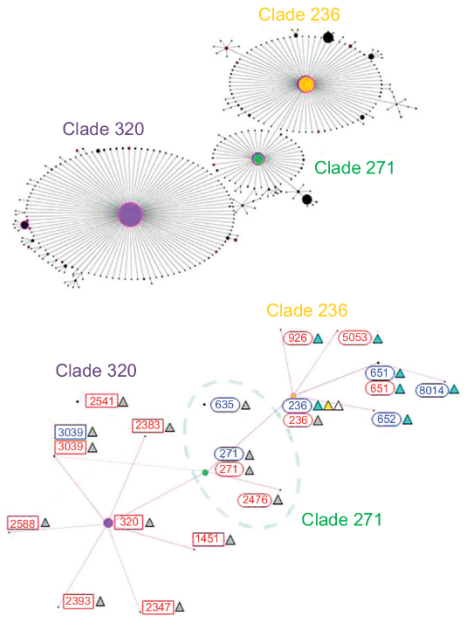


FIGURE 5.1 Comparison of clonal complex 320 (CC320) multilocus sequence types in the Atlanta genome collection with all known CC320 sequence types. Top panel: eBURST analysis of clonal complex CC320 in the spneumoniae.mlst.net database (as of November 2014). CC320 is composed of three subgroups, or clades, each named after the predicted founding sequence type (ST) of the clade. Each dot represents a unique ST with a diameter proportional to the number of representatives in the database. The founding ST is located at the center of each clade. Clade 320 was founded by ST320 (purple), clade 271 was founded by ST271 (green), and clade 236 was founded by ST236 (orange). Lines connecting each ST represent a single locus variation between the two types. Sequence types circled (pink) are represented in the Atlanta genome collection. Bottom panel: eBURST analysis of CC320 in the Atlanta pneumococcal genome collection. Sequence types and clades are color-coded as above. Numbers indicate the ST number of each type. Boxed ST numbers indicate that at least one member of the ST was serotype 19A. Numbers enclosed by a rounded rectangle contain a serotype 19F isolate. ST connected by a pink line were single locus variants. Blue numbers indicate that an isolate of that ST was isolated prior to the introduction of PCV7 in the Atlanta metropolitan area (i.e., prior to November 2000). Red indicates that the ST was isolated post-PCV7. Triangles indicate the presence or absence of mobile elements encoding macrolide resistance: gray, Tn*2010*; light blue, Tn*2009*; yellow, Mega; open, susceptible isolate without macrolide resistance element.

erythromycin. Each resistant isolate contained macrolide efflux genes *mef*(E) and *mel*, encoded on the macrolide efflux genetic assembly (Mega) integrated directly into the pneumococcal chromosome or nested within a Tn*916*-like element, either Tn*2009* or Tn*2010*. Tn*2010* also contained the *erm*-type methylase gene *erm*(B) (Figure 5.1). Interestingly, Tn*2010* was identified only in post-PCV7 isolates in clade 236 (Figure 5.1). The earliest isolated CC320 strain in this genome collection was GA04375, a 19F, ST236 isolate from 1995. GA04375 did not contain the Mega- and *erm*(B)-containing transposon Tn*2010* that is commonly found in ST320. It contained instead the Mega element integrated into the RNA methyltransferase gene (*rumA*) located at the left junction of the pneumococcal pathogenicity island-1 (PPI-1, [12]), which was partially deleted. The genome of GA04375 clustered closely with serotype 19F isolates from 1999, GA13499 (ST236) and 3063-00 (ST652). Strain 3063-00 contained Mega integrated the DNA-3-methyladenine glycosidase gene (TIGR4 annotation, SP_0108) instead of *rumA*. GA13499 was sensitive to macrolides and contained no macrolide resistance determinant. This demonstrated the independent acquisition of macrolide resistance by closely related isolates prior to PCV7 introduction. CC320 strains isolated after 2000 were mostly 19A, and all CC320 19A isolates contained Tn*2010*, suggesting that the clone acquired Tn*2010* prior to the 19F to 19A serotype switch. This is supported by the observation that Tn*2010* is inserted into the same locus and with identical junction sequences within the chromosome, regardless of serotype. The core analyses of CC320 isolates revealed the influence of antibiotic pressure, vaccine pressure, and the passage of time on the evolution of pneumococcal clones. It will be interesting to see if CC320 emerges with a new serotype in the post-PCV13 era.

Clade 271 correlated to the ST271 subgroup of CC320 and included only macrolide-resistant serotype 19F isolates (Figure 5.1). Unlike clade 236, clade 271 isolates all contained Tn*2010*, including those dating to the pre-PCV7 era. Clade 320 correlated to the CC320 subgroup founded by ST320. Clade 320 represents a PCV7 escape clone, containing mechanisms to avoid antibiotic and vaccine pressures (Figure 5.1). All clade 320 isolates were serotype 19A, indicating a serotype switch, and dual-macrolide resistance determinants were encoded on Tn*2010* (Figure 5.2). Tn*916*-like elements harboring macrolide resistance elements were in the Atlanta population prior to the appearance of ST320 in Atlanta, and indeed globally. It is believed that the clone developed in Asia and was subsequently disseminated worldwide. Epidemiology data supports this theory [13]. However, our sampling of CC320 in Atlanta isolates reveals CC320 clonal diversity in Atlanta similar to that observed around the world. This suggests that the ingredients for the ST320 superbug evolution exist in local populations globally and raises the possibility that convergent evolution could result in similar clones developing independently in local pneumococcal populations. Does this mean that under the correct selective pressures, the formation of the superbug, or a similar dominant clone, was inevitable and will happen again? Significantly, while many of the precursors of ST320 were present in Atlanta prior to PCV7, serotype 19A ST320, with its characteristic dual macrolide resistance determinants encoded by Tn*2010*, did not appear in Atlanta until a single isolate was identified post-PCV7 in 2003 (unpublished). The earliest ST320 isolate in the genome study was from 2004 (Figure 5.2).

Because the ST320 clone emerged globally, including in locations with poor or no PCV7 coverage, antibiotic resistance was thought to be the major selective force driving its emergence in these populations. However, this did not explain why 19A displaced multidrug-resistant 19F in non-vaccinated populations. Recent findings have suggested that the ST320 clone was a better colonizer of the nasopharynx than its progenitor ST236 [13]. The increased fitness was not explained by the capsule difference [13]. This suggested that ST236 had acquired, by transformation and recombination, genes involved in colonization as well as the 19A capsule locus. Thus, it appears the evolution of ST320 was a result of multiple selective pressures.

Another geographically restricted study focused on more than 3000 carriage isolates from a 2.4 km^2 refugee camp at the border of Thailand and Myanmar [14]. Hierarchical clustering of genomes based on sequence similarity revealed clusters that roughly corresponded to MLST-based clonal complexes. This study revealed that among the 3085 carriage isolates sequenced, the largest capsule phenotype group (512 isolates) was composed of non-typable *S. pneumoniae*. These non-typable isolates appear to act as a reservoir of recombinant DNA, especially drug resistance determinants, for the different *S. pneumoniae* lineages; and given that these non-typable isolates are not targeted by current polysaccharide vaccines, they will continue to be carried. The study also identified hotspots of recombination within the *S. pneumoniae* genome. These indicate that there are a limited number of genes in which diversity accumulates as a consequence of recombination. It is likely that host and therapeutic pressures underlie this phenomenon. Indeed, the six most prominent hotspots among the refugee camp isolates harbored genes encoding cell surface antigens (*pspA* and *pspC*) and genes associated with resistance to antibiotics (*pbp1a*, *pbp2b*, *pbp2x*, and *folA*).

These applications of whole genome sequencing and analysis to different collections of *S. pneumoniae* illustrated the power of

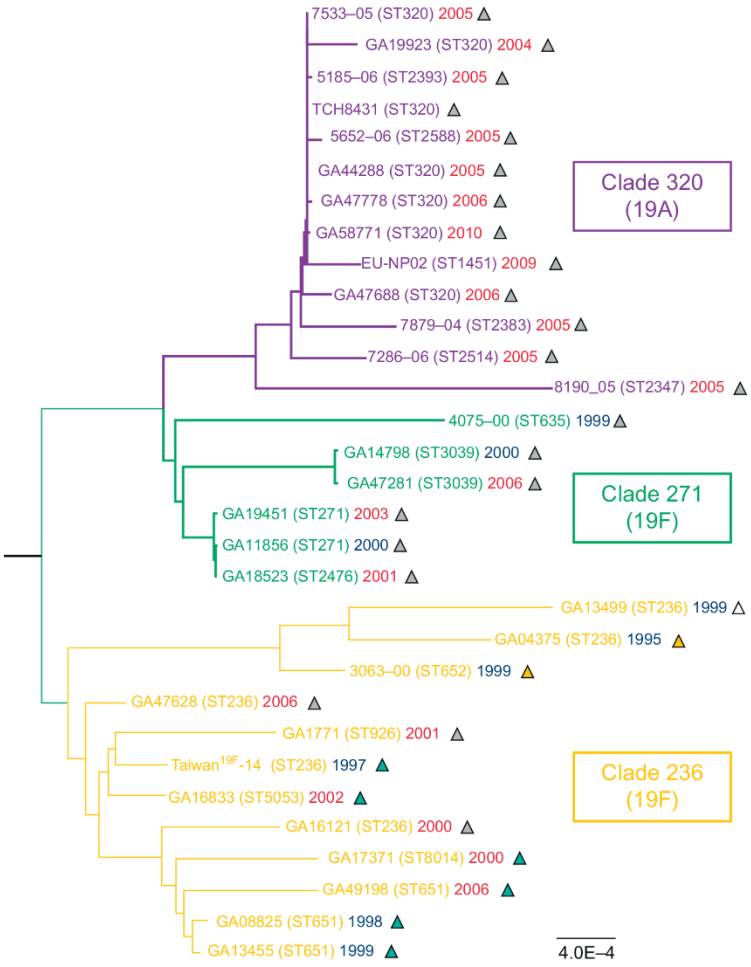


FIGURE 5.2 Core genome phylogeny of CC320 isolates from metropolitan Atlanta. Clades are color-coded as in Figure 5.1. Purple, clade 320; green, clade 271; orange, clade 236. Each clade harbors a single serotype, which is indicated in parentheses within the boxed clade labels. Branch tips of the tree are labeled with strain names, followed by ST (in parentheses) and year of isolation. Years are color-coded to distinguish between strains isolated prior to and after the introduction of PCV7 in the Atlanta metropolitan area (blue and red, respectively). Triangles represent the presence of macrolide resistance elements and are color-coded as in Figure 5.1: gray, Tn*2010*; light blue, Tn*2009*; yellow, Mega; open, susceptible isolate without macrolide resistance determinant.

genomics in providing critical insights into the biology of this species. Current sequencing efforts are focused on filling in the pneumococcal space of genomic diversity, geographical origin, and time (evolution). For instance, the ongoing global pneumococcal sequencing project (GPS, http://news.emory.edu/stories/2013/03/video_pnemonia_genome/) aims to sequence the genome of 20,000 pneumococcal strains isolated before and after the introduction of vaccines in developing countries. The team aims to better characterize vaccine escape and devise next-generation vaccines that avoid that escape.

S. *pneumoniae* Pan-Genome

S. pneumoniae displays extensive genomic diversity. This is reflected in the analysis of its pan-genome, the entire repertoire of genes accessible to the *S. pneumoniae* species, which was determined to be much larger than the genome of any individual strain or isolate [15–18]. In fact, the *S. pneumoniae* pan-genome was defined as open, its size increasing logarithmically, meaning that extrapolation based on the 44 genomes sequenced in 2010 suggested that every new genome sequenced contributed new genes to the species, and the trend indicated that a very large number of genomes would have to be sequenced to fully characterize the entire gene repertoire [15,16].

A pan-genome analysis we performed based on 158 isolates from the Atlanta metropolitan area and other publicly available genomes confirmed the trend based on 44 genomes (Figure 5.3). The new genes power law regression equation for 158 genomes was:

$$y = 269.3229 \pm 2.7959x^{(-0.9821 \pm 0.0028)}$$

This formula allows for extrapolation that can be used to predict the number of new genes that would be identified given increasing numbers of additional genomes sequenced; this is shown in Table 5.1. It suggests that after 500 genomes sequenced, every other genome will provide a new gene on average (~0.5 new gene per genome); after 1000 genomes every third genome will still provide a new gene; and so on. This, of course, depends on the randomness of sampling for strains to be sequenced. Thus, the pan-genome of *S. pneumoniae* is still predicted to be extremely large. This has broad implications for the biology of *S. pneumoniae*. The core genome (shared by all strains) typically includes genes responsible for the basic aspects of the biology of the species and its major phenotypic traits. By contrast, dispensable genes (shared by a subset of the strains) contribute to the species diversity and might encode supplementary biochemical pathways and functions that are not essential for bacterial growth but which confer selective advantages, such as adaptation to different niches, antibiotic resistance, or colonization of a new host [19]. Donati et al. [15] predicted that a fast-growing pan-genome, with strains that are quickly diversifying by integrating new genes, indicates that the species is exploring novel evolutionary possibilities. They postulate that the *S. pneumoniae* species is possibly adapted to its current ecological niche but remains open to the acquisition of new genes while maintaining stability.

Van Tonder et al. [20] used a Bayesian approach to estimate a bacterial core genome that, unlike the classical pan-genome analysis described above, does not require that every single isolate sequenced harbors all core genes. This accommodates for the possible presence of rare strain variants that may be missing some genes that would otherwise be considered core. Application of the model to 336 *S. pneumoniae* genomes encompassing 39 serotypes, 32 countries, and 90 years of isolation estimated the presence of 948 core genes. Another method, based on clusters of orthologous genes, also applied by van Tonder et al. predicted 1194 core genes. Generally speaking, differences in core gene counts arise from the

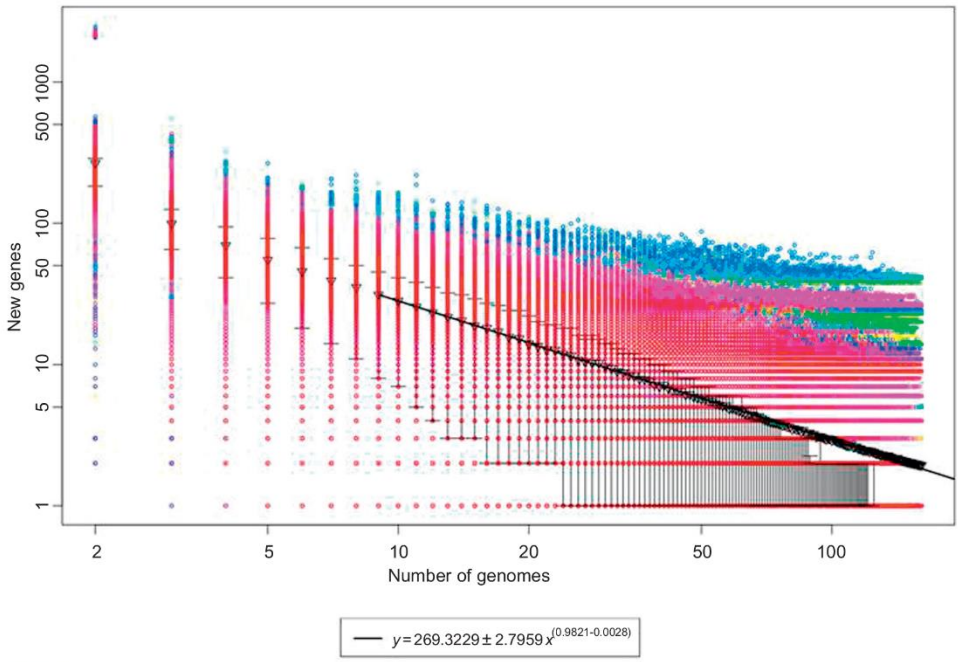


FIGURE 5.3 **New gene discovery graph from the pan-genome analysis of 158 *S. pneumoniae* isolates.** For each reported number of genomes (n), the circles represent the number of new genes found in different randomly chosen combinations. Triangles show the mean values for each distribution. The curve and equation represent a power law regression for new genes discovered that was fitted to the means of new gene counts (triangles) for each value of n.

use of different data sets (the more genomes analyzed, the smaller the core genome), different models, and varying methods for alignment and different cutoffs, including percent identity and whether or not alignment had to occur over the full length of the genes (estimated core genome model) or half of the gene length [17].

Dispensable Genome and Mobile Elements

The concept of what constitutes a pneumococcal clone (or strain, or isolate) has evolved with our ability to look into the genetics behind pneumococcal evolution with increasing resolution. Capsule serotyping was insufficient as a means to infer phylogeny. Several typing methods were improvements, but not until MLST provided a standardized and accessible method whereby isolates could be reliably typed and compared to typed isolates from any location [21]. MLST demonstrated that a single sequence type can exist with many different capsule serotypes due to serotype switching [8,22]. Comparative genomics revealed a surprisingly large amount of variation within the genomes of "identical" clones as defined by MLST.

TABLE 5.1 Use of a Pan-Genome Analysis to Extrapolate the Number of New Genes That Would Be Identified Given Increasing Numbers of Additional Genomes Sequenced

Number of genomes sequenced	Estimated number of new genes identified per genome	Estimated number of genomes needed to find a new gene
158	1.87	0.54
200	1.48	0.68
500	0.60	1.66
1000	0.30	3.28
5000	0.06	15.94
10,000	0.03	31.49
20,000	0.02	62.20
50,000	0.01	152.96

Variability between individuals within a single clonal complex is due primarily to variations in the gene content of their dispensable genomes, that is, those genes not essential for survival and pathogenicity of the pneumococcus. The dispensable genome accounts for approximately one-quarter of a pneumococcal genome [15]. The contents of the dispensable genome of a pneumococcal clone are heavily influenced by horizontal gene transfer of material from other pneumococci, closely related commensal streptococci, and/or more distantly related bacteria. The pneumococcus is extremely adept at acquiring novel DNA through transformation and recombination, allowing virtually unrestricted flow of gene content within pneumococcal populations. Frequent recombination events lead to constant shuffling of the dispensable genome and often to swapping of large regions of the core genomes. This process has created genotypic heterogeneity within pneumococcal populations [22]. Successful clones emerge from the milieu of genotypes due to selective environmental and host-related pressures. Antibiotic pressure selects for genotypes that include antibiotic resistance genes. Vaccine pressure suppresses targeted capsule genotypes, thus promoting the emergence of non-vaccine capsule genotypes. The pneumococcal capsule conjugate vaccines (PCV7, PCV13) selected for non-vaccine capsule serotypes including serotype switch "escape" mutants [8].

Transformation and recombination can also be a genome-stabilizing factor. There is no SOS response system encoded within the core genome of *S. pneumoniae* to repair damaged DNA [23]. Instead, pneumococci repair damaged DNA by allele replacement through recombination with undamaged DNA acquired by transformation [24]. The "sharing" of DNA between pneumococcal cells to repair randomly located DNA lesions can have a homogenizing effect on the pneumococcal genome.

Clonal diversity is also limited by the recombination-dependent phenomenon of soft selective sweeps. Interspecies transformation events have been demonstrated to transfer fragments ranging in size from 0.4 to 235 kb [25]. Transformation of genes providing a selective advantage can be linked to neutral or even detrimental genes. The closer two genes are on the chromosome, the more likely they are to be transferred concomitantly and the stronger the selective pressure on the indirectly selected gene.

In the Atlanta genome collection, soft sweeps driven by macrolide resistance are apparent. Tn*5253*-like conjugative transposons carry the chloramphenicol acetyltransferase gene (*cat*), conferring resistance to chloramphenicol. In all instances, Tn*5253*-like elements were inserted into the ribosome maturation protein *ylqF* (TIGR4 annotation, SP_1154) [26]. Nested in the Tn*5253*-like elements were Tn*916*-like elements carrying the tetracycline resistance gene *tet*(M). Inserted in conserved loci of the Tn*916*-like elements were various macrolide resistance elements including Mega and the *erm*(B)-containing elements Omega and Tn*917*. Macrolide-resistant Tn*916*-like elements associated with Tn*5253*-like elements included Tn*6002* (Omega), Tn*2009* (Mega), Tn*2010* (Omega and Mega), or Tn*3872* (Tn*917*). This indicates that selection for recombination of the Tn*916*-like elements, or fragments thereof, into larger elements allows interconversion between Tn*916* and the macrolide-resistant version of Tn*916*. Selection for these recombination events by macrolide exposure also provides a soft selective sweep for chloramphenicol and tetracycline resistance.

The efficiency of homologous recombination depends upon suitable regions of homology between donor and recipient cells. This becomes a barrier to horizontal gene transfer between distantly related bacteria. Transfer between species lacking extensive homology with pneumococcal chromosomes is facilitated by mobile DNA elements such as insertion sequences, phages, and integrative and conjugative transposons (ICE). Mobile elements are excised, transferred, and integrated in a target sequence–dependent manner, thus bypassing the need for homology between donor and recipient. These elements often carry "cargo" genes, which may come from distantly related bacteria and which may be beneficial to the pneumococcus. Once integrated into a pneumococcal chromosome, the mobile element and its novel gene cargo can be disseminated within the pneumococcal population by transformation and recombination. Thus, mobile elements are a means of expansion of the pneumococcal pan-genome through additions of novel gene content to the accessory genome.

In conclusion, *S. pneumoniae* has a large and growing pan-genome. Horizontal gene transfer is mediated by transformation and recombination, and by the movement of mobile elements. The continual shuffling of gene content within the pan-genome and the occasional acquisition of novel DNA from non-pneumococcal bacteria have led to tremendous variation in the genetic background of pneumococcal clones circulating in a population. Selective forces including antibiotics, vaccines, inter- and intraspecies competition, and host defenses pull the previously existing rare strains, from the milieu of pneumococcal genotypes, that are most suitably adapted to deal with the environmental challenges at hand at any given time. Widespread selective pressure, such as antibiotic usage and vaccination, subsequently promote the clonal propagation and dissemination of successful clones. Future genomic studies will aid in improving our understanding of the relative roles played by genetic variability in local and global pneumococcal populations.

VARIATION AND VIRULENCE

S. pneumoniae (the pneumococcus) colonizes the human nasopharynx and in some cases can cause diseases such otitis media, pneumonia, and meningitis. The pneumococcus produces a range of colonization and virulence factors including a polysaccharide capsule, surface proteins and enzymes, and the cytoplasmic toxin pneumolysin. In terms of ability to cause disease, not all pneumococci are equal. Some strains or serotypes are rarely associated with disease, while others are often associated with invasive disease [27,28]. The ability of pneumococci to cause disease in humans is related to

the genetic content of the organism, such that the presence or absence of virulence genes and/or variation in the sequence of virulence genes dictates the virulence of the strain. Several screens have been conducted to identify genes important in pathogenesis of infection. These include signature-tagged mutagenesis in animal models of infection and colonization [29–32] as well as TnSeq-based screening [33]. Microarray-based studies and whole genome sequencing have been used in attempts to determine the complement of genes required to define the ability to cause invasive disease [27,34,35]. These studies assume that there is an essential core genome and that the differing virulence of pneumococcal strains is determined by a set of accessory genes in the pneumococcal chromosome. It has proved difficult to associate the ability to cause invasive disease with particular genetic loci. The capsule locus is essential for virulence, but not all capsulated strains cause disease. Serotypes associated with the highest rates of invasive disease are 1, 4, and 7F. However, there may also be differences in ability to cause invasive disease among clonal types of the same serotype [27]. Blomberg et al. [27] conclude that the accessory regions required for invasive disease may be redundant as no unique pattern distinguishes the most invasive pneumococcal clones from others. Gene content may also be reflected in different regulatory pathways within strains of pneumococci. In addition to variation in gene content there is also variation in the sequence of individual genes known to be important for virulence, such that SNPs may define the virulence profile of some strains. By understanding the effect of these sequence variations on the ability of pneumococci to cause disease, it may be possible to define more subtle mutations (rather than presence and absence of genes) that allow pneumococci to vary in invasive potential. Some of these key genes and the biological effects of variation in sequence of these virulence factors are considered here.

Capsule

The polysaccharide capsule is the most important virulence factor in the pneumococcus and is the basis for serotyping of pneumococci. There are 94 known serotypes [36–40]. The genes for the biosynthesis of 93 of the capsule types are found in the same location in the pneumococcal chromosome, between the *dexB* and *aliA* genes. The exception to this is serotype 37, which is synthesized from a single gene elsewhere in the chromosome. One of the most striking features of the pneumococcal capsule locus is its huge genetic divergence, as only a few genes are conserved among the different clusters [36,41]. The capsule protects the pneumococcus from phagocytosis [42]. Antibody to cell wall constituents binds to the surface of the pneumococcus and in turn binds complement components. Presence of the capsule prevents iC3b and the Fc of immunoglobulins bound to the bacterium from interacting with their receptors on the surface of phagocytic cells, with the result that the bacteria cannot be taken up and killed by the phagocyte [43]. The capsule is also crucial for colonization as it prevents removal by mucus [44] and can also restrict autolysis and reduce exposure to antibiotics [45]. Pneumococci lacking a polysaccharide capsule can be isolated from the upper respiratory tract of humans [46]. These strains are often referred to as nontypable and are usually associated with asymptomatic carriage but are also associated with outbreaks of conjunctivitis [47] and occasionally with invasive disease [48,49]. Non-typable strains can be divided into two groups. Group I are those with a disrupted or nonfunctional capsule locus and Group II are those isolates that contain genes not found in normal capsular types [50]. Group II can be further divided

into NCC1 and NCC2. NCC1 isolates have the *pspK* gene present at the site of the capsule locus [51]. The *pspK* gene codes for a novel pneumococcal surface protein that may play a role in colonization [52]. The *pspK* gene has also been named novel surface protein gene A (*nspA*) [53]. The *nspA* gene is present along with a variety of intact and disruptive IS elements. The *nspA* gene itself shows high levels of conservation in some areas, with a hypervariable repeat region: no two isolates are identical [53]. NCC2 isolates have both the *aliB*-like ORF1 and *aliB*-like ORF-2 genes [50,51,53]. Analysis of population structure shows that *nspA* is not restricted to a single lineage of closely related pneumococci but is found in distantly related isolates. The presence of this gene in strains isolated from distant geographical locations suggests that strains carrying this gene are successful [53].

Surface Proteins

Analysis of the genome sequence of *S. pneumoniae* strain TIGR4 [3] identified 70 genes for proteins predicted to be exposed at the cell surface. These proteins are surface attached by one of three mechanisms: peptidoglycan anchor motif (LPXTG), choline-binding motif, or lipid-attachment motif [54]. The LPXTG motif allows the enzyme sortase-A to covalently link the protein to the bacterial cell wall by linkage of the threonine of the motif to the pentaglycine linkage of peptidoglycan in the pneumococcal cell wall [55]. The number of LPXTG proteins can differ between strains, and several of these proteins are known to be associated with the virulence of the organism. Key LPXTG-anchored proteins are neuraminidase A (NanA), serine protease PrtA and hyaluronidase.

Neuraminidase cleaves *N*-acetyl neuraminic acid from glycolipids, lipoproteins, and oligosaccharides in host cells, which may unmask binding sites for the organism. NanA plays a role in colonization and development of otitis media in a chinchilla model [56]. Loss of sialic acid as a result of neuraminidase activity accompanies the spread of pneumococci along the eustachian tube to the middle ear [56]. NanA plays an important role in biofilm formation, and sialic acid released by the action of NanA may be an important signal in regulation of pneumococcal virulence [57]. The *nanA* gene is present in all clinical isolates [58–60]. The *nanA* gene shows high sequence diversity that may be important in avoidance of the host immune response [58]. The original cloning of the *nanA* gene from *S. pneumoniae* strain R36a (NCTC 10319) isolated an enzymatically active clone [61]. Subsequent sequence analysis showed this clone was not complete, lacking 233 amino acids from the C-terminus. Interestingly, in the original genome sequencing project of TIGR4 the *nanA* sequence is annotated as a pseudo-gene due to the presence of an 11 base pair deletion that results in a changed reading frame and termination of the gene at amino acid 804 (of a possible 1035) [3]. However, the enzymatic portion of the protein is intact and can be isolated from the pneumococcus. The absence of the C-terminal part of the protein means that the LPXTG anchor is missing and the enzyme is not linked to the cell wall. This may be important in the pathogenesis of disease as NanA is important in binding pneumococci to human cells, including those of the blood–brain barrier [62,63], and lack of surface anchoring may compromise this function.

Hyaluronidase breaks down the hyaluronic acid component of mammalian connective tissue and extracellular matrix and is produced by clinical isolates of pneumococci [64]. The degradation of hyaluronic acid may aid bacterial spread and colonization. Hyaluronidase may also potentiate pulmonary inflammation by complex interaction with chemokines and cytokines. TNFα and IL-1β are able to induce the production of hyaluronic acid by fibroblasts [65], which can then promote further cytokine

secretion by binding to CD44 on host cells. The system is further complicated by the ability of IL-1 to release host hyaluronidase. Breakdown products of hyaluronic acid stimulate chemokine production by macrophages [66], which increases cell recruitment and inflammation. Some serotype 3, ST180 strains contain an SNP at position 376 of the hyaluronidase coding sequence, which results in a stop codon and truncation of the protein after 125 amino acids. These strains produce no active hyaluronidase.

The *prtA* gene has been confirmed in all pneumococcal isolates tested [67]. PrtA is a serine protease and is required for full virulence in animal models; vaccination with the protein provides protection from infectious challenge [67]. PrtA plays a role in the killing of *S. pneumoniae* by apolactoferin [68]. Expression of *prtA* is co-regulated with a number of other virulence genes, including those encoding the pilus and pneumolysin genes, by the transcriptional factor PsaR [69]. Regulation of *prtA* expression by PsaR has also been demonstrated to be oppositely repressed and stimulated by manganese or zinc [70].

Three proteins from TIGR4 have LPXTG-like motifs; these are the pilin proteins (SP_0462, SP_0463, and SP_0464), which are linked to each other by specific pilus-sortase enzymes [71]. These genes are part of the *rlrA* pathogenicity islet [72] and are transcribed together on the same transcript by the adjacent transcriptional regulator. The *rlrA* pathogenicity islet codes for the production of the pneumococcal pilus; this genetic locus is present in less than 20% of clinical strains [73]. A second pilus type, which is involved in adherence of pneumococci to epithelial cells, has been identified, and some strains can express both types [74].

Choline-binding proteins (CBPs) are anchored to the cell surface via the interaction of repeat domains of the protein with choline residues present in the pneumococcal cell wall. Teichoic and lipoteichoic acids in the cell wall are decorated with phosphorylcholine residues that anchor the CBPs to the pneumococcal cell. These proteins have repeated sequences of approximately 20 amino acids (choline-binding module), usually present in the C-terminal region of the protein. Two to twelve modules form the choline-binding domain that is attached to phosphorylcholine in the cell wall in a noncovalent manner. CBPs may have various enzymatic activities or may have binding properties to allow binding to host cells or extracellular matrix [54]. Analysis of the genome sequences of pneumococcal strains R6 [75] and TIGR4 [3] predict 12 CBPs in R6 and 15 in TIGR4 [76]. Several CBPs are associated with the ability to bind to host proteins.

The genome of *S. pneumoniae* contains approximately 40 genes predicted to code for lipoproteins [3,54], many of which are involved in virulence as part of nutrient uptake transporters. Cation ABC transporters have major effects on pneumococcal virulence, with loss of PsaA manganese transporter lipoprotein or combined loss of AdcA and AdcAII zinc or the PiaA and PiuA iron ABC transporter lipoproteins, resulting in strains of greatly reduced virulence [77–82]. The mechanism of lipoprotein attachment to the bacterial cell membrane and processing is conserved among bacteria. Prolipoproteins are secreted by the general secretory pathway and then are covalently linked to the cell membrane by the enzyme diacylglyceryl transferase (Lgt) [82]. A type II lipoprotein signal peptidase (Lsp) then cleaves the N-terminal signal peptide adjacent to the "lipobox" cysteine residue to form the mature lipoprotein [83]. Deletion of the *lgt* gene from pneumococcus has widespread effects on ABC transporter functions that collectively prevent the mutant from establishing invasive infection [84].

Pneumolysin

Pneumolysin (Ply) is a 53-kDa pore-forming toxin made by almost all clinical

isolates of the pneumococcus; it is expressed during the log phase of growth [85]. There are at least 16 different naturally occurring variants of Ply including allele 5 Ply, which is expressed in specific strains of serotypes 1 and 8 pneumococci [86–88]. It has been demonstrated that both human and murine mononuclear cells exposed to *S. pneumoniae* that express fully lytic toxin produce IL-1β, and the production of this cytokine depends on the NOD-like receptor family, pyrin domain containing 3 (NALP3) inflammasome [89]. Strains expressing the nonhemolytic allele 5 of the toxin did not stimulate IL-1β production. NLRP3 activation was beneficial for mice during pneumonia caused by pneumococcal strains expressing fully active toxin due to cytokine production and maintenance of the pulmonary microvascular barrier. Thus, polymorphisms in the pneumolysin protein may substantially affect recognition of bacteria by the innate immune system. Pneumolysin is produced by virtually all clinical isolates of the pneumococcus [90], and analysis of the pneumolysin gene sequence from 121 clinical isolates identified 14 protein alleles [87], some of which are associated with lack of hemolytic activity of the toxin. Some clinical strains were shown to have insertions of either a section of duplicated sequence or transposon IS1515 [87,91], suggesting that pneumolysin is not absolutely essential for the pneumococcus to be able to cause infection. Although thought to be released only when pneumococci undergo autolysis [92], Ply can be released independently of the major autolysin [93]. Ply has been shown to be exported to the cell wall [94,95]. Ply plays several roles in infection. The toxin appears to have no role in inflammation associated with meningitis [96–98] but does have a role in deafness associated with meningitis [98] and in bacteremia [99] and pneumonia [100]. Ply has been suggested to play a role in damage to the blood brain barrier as it his responsible for the majority of cytotoxicity in brain microvascular endothelial cells exposed to *S. pneumoniae in vitro* [101].

Two-Component Systems and Regulation of Virulence

Bacterial adaptation to the external environment is often mediated by two-component systems (TCS). A typical TCS is composed of a membrane-bound sensor histidine protein kinase (HK) and a cognate response regulator (RR), which is usually a DN-binding protein. On stimulation by an appropriate signal the HK is auto-phosphorylated on a conserved histidine. The phosphate group is then relayed to an aspartate residue in the RR. The availability of pneumococcal genome sequences reveals 13 HK:RR pairs and a single "orphan" RR with no associated HK [102,103]. Genetic studies have been conducted to define the roles of these systems in virulence. Lange et al. [102] analyzed the effect of gene deletions on the virulence of serotype 3 and serotype 22 pneumococci in a mouse model of systemic infection and found no effect on virulence. Throup et al. [103] used a mouse model of pneumonia and demonstrated a role in infection for most of the TCS in a serotype 3 strain (0100993). Thus the role of TCS in virulence is dependent on the genetic background of the bacterial strain as well as route of infection. This finding was confirmed by Blue and Mitchell [104], who found that deletion of the TCS09 system had no effect on the virulence of strain D39 in murine models of pneumonia and bacteremia; the same mutation in strain 0100993 caused attenuation in the pneumonia model but not in the systemic disease model. There are many other examples of different effects of regulatory genes on virulence depending on the strain of pneumococcus studied. It is becoming increasingly clear that to understand these regulatory processes we need to study the control of bacterial gene

expression in the *in vivo* environment as well as the detailed regulatory pathways involved in the processes of bacterial colonization and development of disease in the host. The use of RNA sequencing to analyze bacterial transcriptomics will allow these detailed relationships to be dissected.

Pneumococcal Strain Variation During Infection

The pneumococcus can undergo genetic changes during the course of an infection. Two isolates of a serotype 3, ST180 were taken from a patient with pneumococcal meningitis [105]. One isolate was grown from the blood and the other from a sample of cerebrospinal fluid (CSF). The two strains were compared by microarray and RNA sequencing, which showed that they had different expression profiles, including a marked up-regulation in the expression of the PatAB transporter in the strain isolated from CSF. Whole genome sequencing identified an SNP present in the regulatory region of the PatAB gene that was associated with changes in gene expression. When the two strains were compared in an animal model of disease, they showed different profiles, with the strain isolated from human blood growing better in mouse blood, while the strain from CSF grew less well in blood but reached higher numbers in the brain. Thus, one SNP can have a marked effect on the virulence of *S. pneumoniae*. The role of recombination also plays a key role in the evolution of pneumococci. Croucher et al. [4] used high-throughput sequencing to analyze 240 isolates of the PMEN-1 ($Spain^{23F}$-1) strain, and more than 700 recombination events were detected, which frequently affected major antigens, including 10 capsule switch events, one of which accompanied a population shift as vaccine escape serotype 19A isolates emerged in the United States after the introduction of the conjugate vaccine. The evolution of antibiotic resistance was observed to occur on multiple occasions. The study shows how genomic plasticity within the pneumococcus can permit adaptation to clinical interventions on very short timescales.

The environment within patients infected with pneumococcus can vary not only according to clinical interventions (antibiotics, etc.) but also due to underlying conditions. For example, children with sickle cell disease (SCD) have a 600-fold increased risk of invasive pneumococcal disease [106]. The increased risk of infection is due to functional asplenia and complement deficiency, and patients also have altered plasma levels of zinc, iron, amino acids, and carbohydrates [107,108]. These patients are routinely vaccinated and prescribed prophylactic antibiotics. Analysis of strains from carriage and disease in the general population and those isolated from patients with SCD shows that strains from SCD patients have specific adaptations [109]. As well as the expected adaptation to antibiotics and vaccination, strains from SCD had undergone gene loss and intragenic recombination to produce mosaic genes. These events had occurred in four key groups of genes responsible for penicillin resistance, capsule biosynthesis, metabolic pathways, and metal ion uptake. The mutations in genes involved in metal ion uptake suggest that these mutations are beneficial in the SCD host and can only be tolerated in this host environment but not in the normal host. Use of a library of Tn-seq mutants in wild type and SCD mice identified genes involved with aspects of SCD pathophysiology in humans, such as abnormal iron homeostasis, purine metabolism, and complement function. One of the six genes identified by Tn-seq analysis is involved in iron uptake into the pneumococcus. The iron transport complex is immunogenic, and loss of this protein may be advantageous in avoiding the immune response. The iron transport process is probably not required or may

even be detrimental to the bacterium in an iron-rich SCD host. The ability of the pneumococcus to thrive despite the loss of antigenic proteins could compromise protective immunity in specific host environments and could influence targets for new protein-based vaccination in SCD patients. The study of Carter et al. [109] also highlights how analysis of bacterial genome sequences from particular disease groups may yield information on the selective conditions within patients suffering from the disease (in this case SCD).

S. PNEUMONIAE AND CLOSE RELATIVES

The pathogenic potential of *S. pneumoniae* is intriguing in the light of its closest relatives, such as *S. mitis* and *S. oralis*, that are rarely associated with disease and are representatives of the commensal microbiota of the upper respiratory tract of humans. Recently, *S. pseudopneumoniae* has been added as a new species, a group of bacteria previously referred to as atypical pneumococci [110–112]. These species are part of the Mitis group of streptococci [113–115]. They are naturally competent for genetic transformation; that is, they can take up DNA and incorporate it into their genome via homologous recombination. It is this property that is the major driving force for genomic diversity within a single species, resulting in a large accessory genome. The molecular mechanism of competence development, a quorum-sensing process based on secretion of the competence-stimulating peptide CSP and its recognition by the TCS ComCD, which is present in all genomes of the above-mentioned species, is well understood (for review, see [116]). However, it is not known under which *in vivo* conditions competence develops, and how frequent gene transfer occurs within and between species. Genomic comparison revealed that in several *S. mitis* and *S. oralis* strains, not all of the 22 essential competence genes described in *S. pneumoniae* are functional or are even absent [117], an indication of fighting genome instability. Nevertheless, *S. mitis* NCTC10712 and *S. oralis* Uo5, which are included in this list, are transformable under laboratory conditions, and it should be noted also that most clinical isolates of *S. pneumoniae* do not develop competence under laboratory conditions, although they contain the entire equipment for competence.

S. pneumoniae inhabits the nasopharynx, whereas oral streptococci reside in the oral cavity. However, signs of interspecies gene transfer, which are obvious from genomic data, indicate that at least occasionally DNA from other species is available. The fact that pneumococcal genomes are significantly larger, approximately 2.1 Mb compared to those of most *S. mitis* or *S. oralis* genomes (~2 Mb), documents a larger accessory genome. Genetic transformation under laboratory conditions results in considerable sequence exchange. After four successive steps of transformation, over 3% of the recipient *S. pneumoniae* genome was replaced by donor *S. mitis* DNA [118]. The recombination events resulted in deletion of one gene, the replacement of a functional gene copy that was fragmented in the recipient strain, and the acquisition of genes not present in the pneumococcal population. The 36 recombination events, spanning between approximately 100 nucleotides up to over 10 kb, clustered in 16 regions throughout the genome. Apparently, gene transfer from *S. mitis* to *S. pneumoniae* also occurs under natural conditions. Phylogenetic trees obtained from all predicted genes among 35 *Streptococcus* spp. revealed clustering of *S. pneumoniae* genes among *S. mitis* genes, which was interpreted as gene transfer from *S. mitis* to *S. pneumoniae* [117]. The size of the regions affected spanned between 116 and 10,600 bp, in agreement with the results reported in the *in vitro* experiment [118].

S. pseudopneumoniae, *S. mitis*, and *S. oralis* are the closest relatives of *S. pneumoniae*. The

finished genomes of members of *S. mitis* B6 [119] and *S. oralis* Uo5 [120] are preferentially used in the following analysis. Despite the relative close relatedness of these species, the overall arrangement of the *S. pneumoniae* genome reveals a striking arrangement, termed X-alignment, when compared to *S. mitis* B6 or *S. oralis* Uo5 (Figure 5.4). Throughout the genome, sequences that are symmetrically inverted with respect to the position of the replication origin or terminus alternate with those that have the same positioning. It is not clear what causes this phenomenon. It has been suggested that inversions might be linked to the replication or termination processes [122]. In *S. mitis* B6, several breakpoints are associated with insertion elements ISSmi1, but the relevance of this observation is not clear [119].

Genomic Comparison—An Overall View

A clear distinction among *S. pneumoniae*, *S. mitis*, and *S. oralis* (Figure 5.5) is obtained by MLST [114,115], which has become the gold standard for analyzing pathogenic bacteria [124] and which is based on sequence comparison of housekeeping genes. MLSA data derived from *S. pseudopneumoniae* strains places this species between *S. pneumoniae* and *S. mitis* [111], but according to MLST analysis the type strain of *S. pseudopneumoniae* as well as strain IS7493, the genome of which is available [125], are found among the *S. mitis* group (Figure 5.5). More genomes of *S. pseudopneumoniae* will be required to clarify the phylogenetic relationship on a genomic basis.

Genomic hybridization of oral streptococci using oligonucleotides based on the *S. pneumoniae* R6 and TIGR4 as well as *S. mitis* B6 sequences revealed an almost smooth transition between these species (Figure 5.6). The explanation is a large accessory genome that circulates among these species but which becomes apparent only by whole genome analysis [126,127] and not by using individual genes that are part of the common core genome.

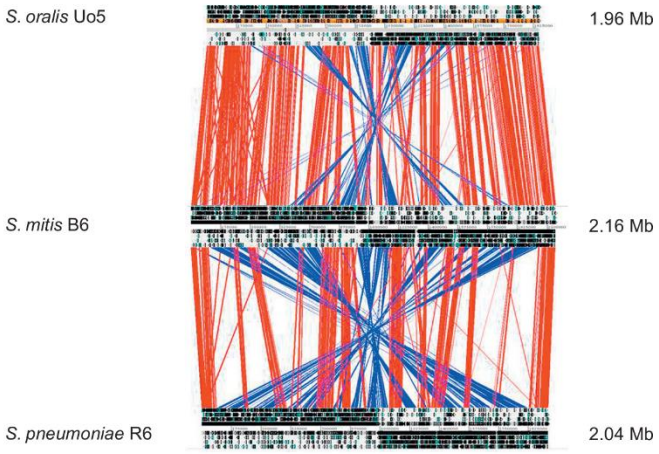


FIGURE 5.4 **Genome alignments of the *S. pneumoniae* R6 genome with those of *S. oralis* Uo5 and *S. mitis* B6.** The alignments are displayed using the ACT program [121]. Red areas mark regions of the same orientation in both species, blue indicates regions implicated in the X-alignment. Only regions greater than 1 kb are shown.

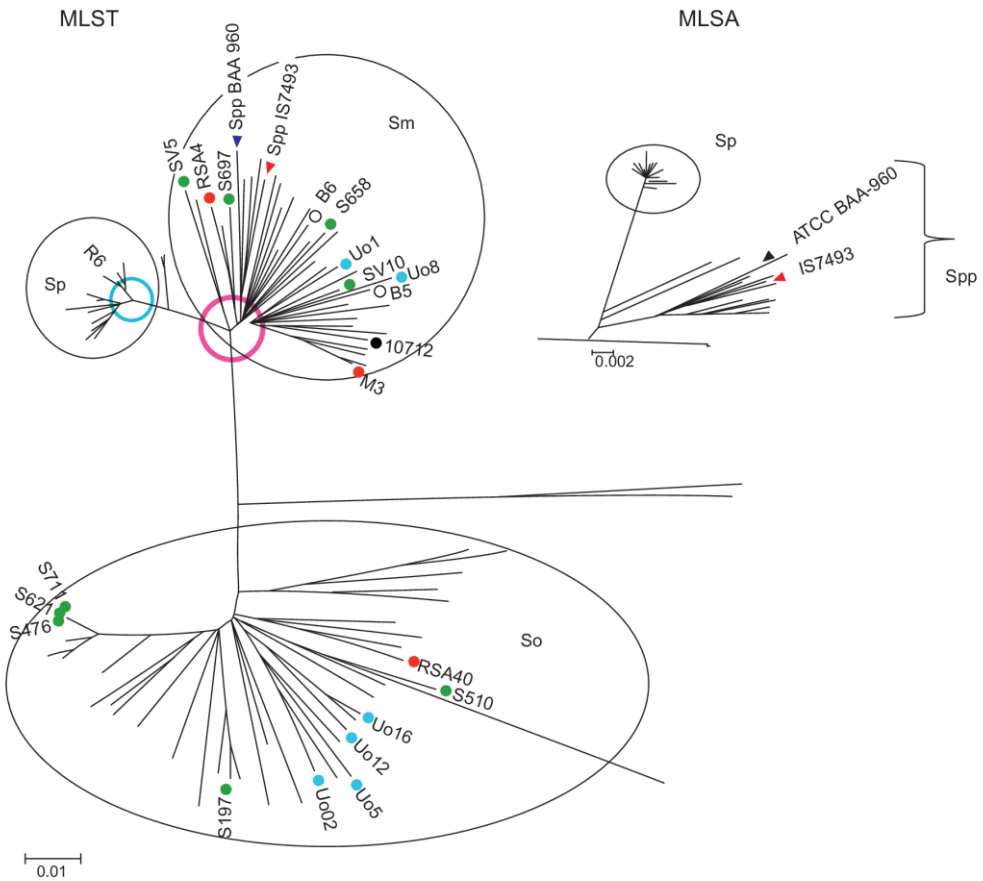


FIGURE 5.5 **Phylogenetic tree of *Streptococcus* spp.** Left: The concatenated sequences of loci used for MLST of *Streptococcus* spp. [115] except that of *ddl* were used for tree construction with the MEGA3.1 program [123]. Sequences were treated as protein coding data. The bootstrap test of phylogeny was chosen as the principal phylogenetic analysis method with the minimum evolution algorithm applied. For bootstrapping the default option of 1050 replicate calculations was chosen. For all other parameters the default options given by the program were used. The color of the dots marks the origin of the strains used in the hybridization experiment shown in Figure 5.6. Red, South Africa; green, Spain; white, Germany; light blue, Hungary; black, reference strains *S. pneumoniae* R6 and *S. mitis* NCTC10712. Sp: *S. pneumoniae*; Sm: *S. mitis*; So: *S. oralis*. For comparison, the type strain of *S. pseudopneumoniae* ATCC BAA-960 and *S. pseudopneumoniae* IS7493 whose genomes are available are included (triangles). Right: multilocus sequence analysis, MLSA [113]. The bracket shows *S. pseudopneumoniae*; Sp: *S. pneumoniae*; others: uncertain. The triangles mark the two *S. pseudopneumoniae* strains used in the MLST tree on the left.

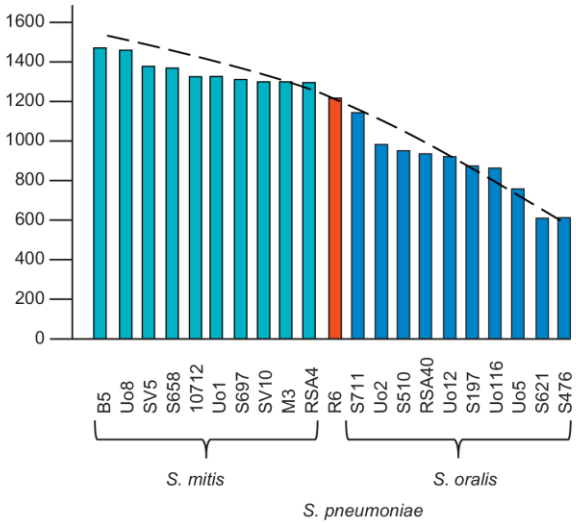


FIGURE 5.6 **Genomic hybridization analysis of *Streptococcus* spp. using a *S. mitis* B6-specific oligonucleotide microarray.** Phage-related gene clusters and mobile elements were not considered. The number of genes giving positive hybridization signals is indicated. The microarray data were evaluated as described [119].

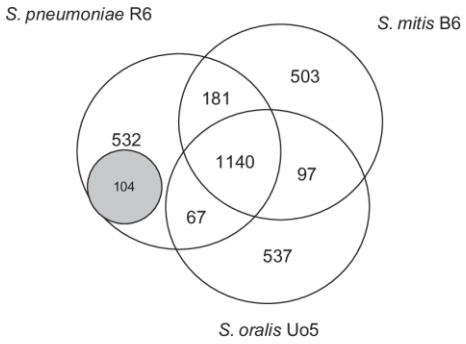


FIGURE 5.7 **Core genomes of *S. pneumoniae* R6, *S. mitis* B6, and *S. oralis* Uo5.** The numbers represent proteins that are 60% identical with a coverage of 70%. When all whole genomes listed in the NCBI microbial genome data base are included, the *S. pneumoniae*-specific proteins drop to 104 (gray circle).

The core genome derived from three strains—*S. pneumoniae* R6, *S. mitis* B6 and *S. oralis* Uo5—includes approximately 60% of the deduced proteins (Figure 5.7), and between 461 and 537 protein encoding genes represent the accessory genome specific to each strain. When the genome of *S. pseudopneumoniae* IS7493 is included, the number of common genes drops to 1105. This number is drastically reduced if more strains of one species are used. When all 26 complete *S. pneumoniae* genomes listed in the NCBI microbial genome database are included, only 104 proteins remain specifically associated with *S. pneumoniae*, with 72 genes being associated with 15 clusters of 2–12 genes. Among them are mostly genes encoding for sugar uptake and utilization systems that are probably a reflection of the special ecological niche

acquired by this species, in addition to components related to its pathogenicity potential, as outlined in the next section. It is clear that with the growing number of genomes all these numbers will need to be adjusted over time.

During the evolution of *S. pneumoniae*, the MLST tree points to two crucial events. One (red circle in Figure 5.5) reveals a common ancestor of both *S. mitis* and *S. pneumoniae*, with *S. pneumoniae* representing one lineage in a cluster of *S. mitis* strains. In fact, each *S. mitis* is approximately as distantly related from each other as from *S. pneumoniae*, and the problem of defining species is obvious. Diversification within the *S. pneumoniae* lineage (blue circle in Figure 5.5) occurred later. It is possible that this second process also reflects the growing population of humans, suggesting that *S. mitis* or *S. oralis* had evolved already in primates. In this context it is curious that specialized serotype 3 clones of *S. pneumoniae* were found in diseased wild chimpanzees in the Thaï National Park, Ivory Coast, which were distinct from human isolates described so far [128,129]. Moreover, *S. oralis* and *S. mitis* could be isolated from primates held in captivity (own unpublished results), suggesting that these oral streptococci might have evolved before *S. pneumoniae* had conquered humans as their optimal host. A curious example of host expansion is the occurrence of type 3 pneumococci that have lost some virulence-associated genes in racing horses [130].

It has been proposed that the three species *S. mitis*, *S. pneumoniae*, and *S. pseudopneumoniae*, arose from an ancient bacterial population that included all *S. pneumoniae*-specific genes [114]. This model was supported recently by genomic analysis of 35 *Streptococcus* spp. Genomic comparison revealed that the average genetic distance from the type strain *S. oralis* ATCC35037 is slightly but significantly larger for *S. pneumoniae* than for *S. mitis*, indicating that the common ancestor was a pneumococcus-like species [117]. The authors provided evidence that interspecies gene transfer occurred mainly unidirectionally from *S. mitis* to *S. pneumoniae*. This includes the import of genes involved in capsular biosynthesis from different groups of streptococci [117], an explanation of the astounding biochemical diversity of the capsule. Similarly, mosaic genes encoding penicillin target enzymes (penicillin-binding proteins) that occur in penicillin-resistant *S. pneumoniae* include sequences that are found in *S. mitis* and *S. oralis* [131,132]. These blocks are larger in *S. mitis* compared to *S. pneumoniae* strains, indicating that they evolved in sensitive *S. mitis* prior to being transferred *to S. pneumoniae* [115]. On the other hand, more genes are decayed in the genome of *S. pneumoniae* R6 or TIGR4 (48 and 62) versus 20 in *S. mitis* B6, excluding IS elements [119], some of them affecting important functions including amino acid biosynthesis, as shown in Figure 5.8. It has been suggested that this functional reduction signifies a "route of no return," that is, fixes *S. pneumoniae* into a current pathogenic lifestyle [119].

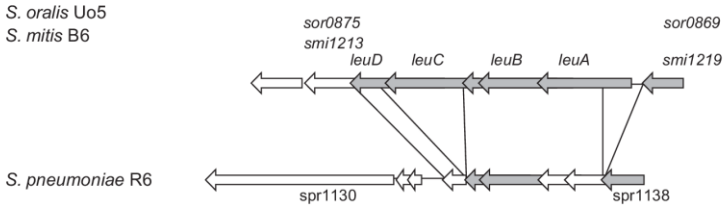


FIGURE 5.8 **Decay of the leucine gene cluster in *S. pneumoniae*.** The leucine gene cluster of *S. mitis* B6 and *S. oralis* Uo5 is shown on top; gray: intact genes.

Virulence Genes of *S. pneumoniae* in *S. mitis* and *S. oralis*

One of the key questions is: Why are oral streptococci not pathogenic; that is, what is specific for *S. pneumoniae* to make it a pathogen?

The choline-containing wall and lipo-teichoic acid (WTA and LTA) have long been believed to be specific features of *S. pneumoniae*. The genes involved in biosynthesis of TA molecules are well conserved in *S. pseudopneumoniae* [125] as well as in *S. mitis*, whereas *S. oralis* strains contain a different cluster, indicating a different TA repeat structure, which is also present in *S. mitis* M3 [133]. All these species also contain CBPs which are associated with TAs by hydrophobic interaction. However, their number is highly variable and differs even among strains of the same species, including *S. pneumoniae* [134]. *S. mitis* B6, with 22 CBPs, represents an unusual example of gene expansion and diversification through gene duplication and recombination events [119]. In contrast, only six CBPs are found in *S. oralis* Uo5, including those that play a principal role in cell physiology: *lytB*, *lytC*, *cbpF*, two paralogues of *cbpD*, in addition to a CBP of unusual repeat structure at the position of *spr0583*/SP_0666, suggesting that they represent the minimum set of physiologically relevant CBPs, and that expansion of CBPs has taken place later in evolution.

The number of LPXTG cell surface proteins that frequently contain repeat motifs predicted to be arranged in coiled-coil structures [119] varies largely not only between species but also within a species. For example, 12 LPXTG proteins are annotated in *S. pneumoniae* R6, 18 in *S. mitis* B6, 17 in *S. pseudopneumoniae* IS7493, and 20 in *S. oralis* Uo5. Many LPXTG proteins of *S. pneumoniae* are found in close relatives as well, the commensal species *S. mitis* [119] and *S. pseudopneumoniae* [111], strongly suggesting that they are important in these commensal species for colonization and interaction with host cells. Also, the *S. oralis* Uo5 genome contains a large number of *S. pneumoniae* homologues [135], similar to other *S. oralis* strains, as suggested from genomic hybridization on a special microarray covering cell-surface proteins and other virulence factors of *S. pneumoniae* R6/TIGR4 and *S. mitis* B6 [135]. The number of homologues detected by microarray analysis, however, represents only a minimal number due to sequence variation of the gene and insufficient coverage by the oligonucleotides. According to genomic analysis, only three LPXTG proteins are common to the four species: the pullulanase gene *pulA* (*spr0247*), an endo-beta-*N*-acetylglucosaminidase (*spr0440*), and an LPXTG protein of unknown function and of different length depending on the number of repeats (*spr0075*). It is curious that LPXTG protein-encoding genes are frequently found in tandem or in close vicinity, indicating either hotspots of recombination or diversification after duplication.

One special example of genome expansion by interspecies gene transfer is the huge serine-rich protein (named MonX, *mon*ster, in *S. mitis* B6 and PsrP, pneumococcal serine-rich repeat protein, in *S. pneumoniae*) and associated genes encoding components involved in export and glycosylation. Serine-rich proteins are common among Gram-positive bacteria [136], but in the Mitis group highly similar clusters occur that differ mainly in the number of glycosyltransferases in the center of the cluster which is greater than 25 kb (Figure 5.9). The reported length of MonX/PsrP varies from 2100 amino acids (aa) to over 4700 aa. However, due to the repeat sequences, the assembly of genome sequencing data is problematic, and Southern blots may be required to confirm its size. In *S. gordonii* it has been described as a platelet-binding protein probably important for oral colonization [137].

There are several other examples of exceptionally large islands specifically associated with only a few strains in one or more species. The *cylM* island (14 kb), encoding for a cytolysin, is

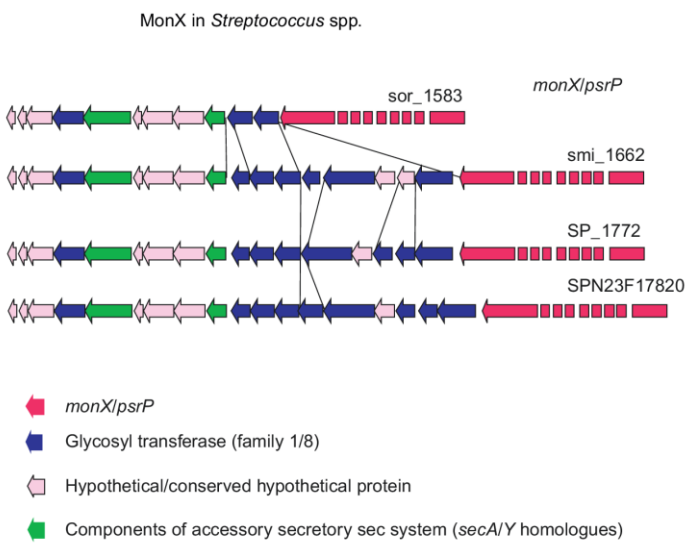


FIGURE 5.9 **The *monX/psrP* cluster in *Streptococcus* spp.** Red, *monX/psrP*; blue, glycosyltransferases; green, components of the accessory sec system (*secA/Y*); pink, hypothetical proteins. Lines indicate regions of similarity.

found exclusively in four *S. pneumoniae* strains. Related genes are common among *Enterococcus*, but not found in other *Streptococcus* spp. Unique to *S. oralis* Uo5 are genes associated with the ESA6 secretion pathway (>45 kb), which is common in mycobacteria [138]. Another island (26 kb), which includes components of the Vtype ATPase, is common among different *Streptococcus* spp. including *S. pneumoniae* TIGR4 [139]. Important in view of antibiotics resistance are Tn*916*-like elements containing the tetracycline resistance determinant *tetM* as mentioned in the previous section, and which occasionally includes erythromycin resistance genes as well (Figure 5.10). *TetM* in *S. mitis* B6 is located on Tn*5801*, which is almost identical to the one described in *S. aureus*, a rare example of interspecies gene transfer between these two species.

Taken together, only a few genes and genomic islands appear to be specifically associated with *S. pneumoniae* in addition to the highly variable capsule cluster: the pneumolysin-autolysin *ply-lytA* island, the CBPs *pspA*, *pcpA*, *pspC* and its variant *hic*, together with the two-component regulatory system TCS06, and the hyaluronidase *hlyA*. In fact, no hyaluronidase activity has been found in *S. mitis* strains [114], but it is present in *S. oralis*. *S. pseudopneumoniae* IS7493 also harbors *ply* and *lytA* in close proximity, but genes in between differ completely from the *S. pneumoniae* island. Occasional isolates of *S. mitis* harbor the Ply gene [114,119,140,141], and the analysis of some *ply*-containing *S. mitis* strains again revealed a genomic environment distinct from that in *S. pneumoniae* [127]. The presence of *hlyA* together with the expansion of sugar-utilizing systems in *S. pneumoniae* might be linked to the conquest of a special ecological niche.

ICE—TetM resistance in *Streptococcus* spp.

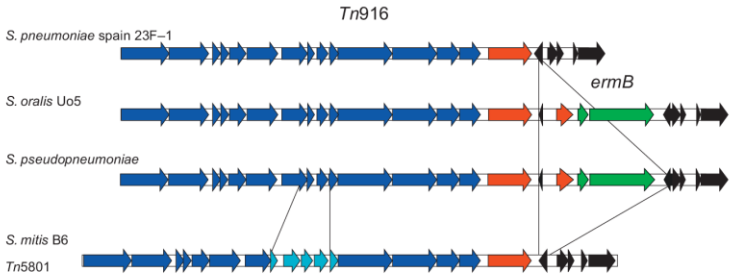


FIGURE 5.10 **ICE elements carrying the tetracycline resistance gene *tetM* in *Streptococcus* spp.** Red, *tetM*; green, *erm* (B); light blue, *S. mitis* B6-specific.

Acknowledgments

This work was supported by a grant to RH from the Deutsche Forschungsgemeinschaft HA 1011/13-1 and HA1011/11-2, and with federal funds to HT from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number HHSN272200900009C (Claire Fraser, PI). Work in the Mitchell Laboratory is supported by grants from the Medical Research Council and Wellcome Trust. Invasive pneumococcal isolates for genome sequencing were provided by the Georgia Emerging Infections Program.

References

[1] McAdam PR, Richardson EJ, Fitzgerald JR. High-throughput sequencing for the study of bacterial pathogen biology. Curr Opin Microbiol 2014;19:106–13.
[2] Dopazo J, Mendoza A, Herrero J, Caldara F, Humbert Y, Friedli L, et al. Annotated draft genomic sequence from a *Streptococcus pneumoniae* type 19F clinical isolate. Microb Drug Resist 2001;7:99–125.
[3] Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 2001;293:498–506.
[4] Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der LM, et al. Rapid pneumococcal evolution in response to clinical interventions. Science 2011;331:430–4.
[5] Lee AW, Tettelin H, Chancey S. Genomic analyses of clonal isolates provide clues to the evolution of *Streptococcus pneumoniae*. Front Microbiol 2011;2:63.
[6] Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von GA, Linares J, et al. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. Genome Biol 2012;13:R103.
[7] Choi EH, Kim SH, Eun BW, Kim SJ, Kim NH, Lee J, et al. *Streptococcus pneumoniae* serotype 19A in children, South Korea. Emerg Infect Dis 2008;14:275–81.
[8] Brueggemann AB, Pai R, Crook DW, Beall B. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. PLoS Pathog 2007;3:e168.
[9] Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 2011;27:334–42.
[10] Sahl JW, Matalka MN, Rasko DA. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl Environ Microbiol 2012;78:4884–92.
[11] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013;30:2725–9.
[12] Brown JS, Gilliland SM, Holden DW. A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. Mol Microbiol 2001;40:572–85.
[13] Hsieh YC, Lin TL, Chang KY, Huang YC, Chen CJ, Lin TY, et al. Expansion and evolution of *Streptococcus pneumoniae* serotype 19A ST320 clone as compared to its ancestral clone, Taiwan19F-14 (ST236). J Infect Dis 2013;208:203–10.
[14] Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 2014;46:305–9.

[15] Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 2010;11:R107.

[16] Muzzi A, Donati C. Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae*. Int J Med Microbiol 2011;301:619–22.

[17] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward L, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 2005;102:13950–5.

[18] Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 2008;11:472–7.

[19] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005;15:589–94.

[20] van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, et al. Defining the estimated core genome of bacterial populations using a Bayesian decision model. PLoS Comput Biol 2014;10:e1003788.

[21] Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology 1998;144:3049–60.

[22] Crisafulli G, Guidotti S, Muzzi A, Torricelli G, Moschioni M, Masignani V, et al. An extended multilocus molecular typing schema for *Streptococcus pneumoniae* demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. Infect Genet Evol 2013;13:151–61.

[23] Prudhomme M, Attaiech L, Sanchez G, Martin B, Claverys JP. Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*. Science 2006;313:89–92.

[24] Claverys JP, Prudhomme M, Martin B. Induction of competence regulons as a general response to stress in gram-positive bacteria. Annu Rev Microbiol 2006;60:451–75.

[25] Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, et al. Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog 2010;6:e1001108.

[26] Chancey S, Agrawal S, Schroeder MR, Farley MM, Tettelin H, Stephens DS. Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. Front Microbiol 2015;6.

[27] Blomberg C, Dagerhamn J, Dahlberg S, Browall S, Fernebro J, Albiger B, et al. Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. J Infect Dis 2009;199:1032–42.

[28] Inverarity D, Lamb K, Diggle M, Robertson C, Greenhalgh D, Mitchell TJ, et al. Death or survival from invasive pneumococcal disease in Scotland: associations with serogroups and multilocus sequence types. J Med Microbiol 2011;60:793–802.

[29] Hava DL, Camilli A. Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. Mol Microbiol 2002;45:1389–406.

[30] Polissi A, Pontiggia A, Feger G, Altieri M, Mottl H, Ferrari L, et al. Large-scale identification of virulence genes from *Streptococcus pneumoniae*. Infect Immun 1998;66:5620–9.

[31] Lau GW, Haataja S, Lonetto M, Kensit SE, Marra A, Bryant AP, et al. A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence. Mol Microbiol 2001;40:555–71.

[32] Chen H, Ma Y, Yang J, O'Brien CJ, Lee SL, Mazurkiewicz JE, et al. Genetic requirement for pneumococcal ear infection. PLoS One 2008;3:e2950.

[33] van OT, Camilli A. A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome Res 2012;22:2541–51.

[34] Obert C, Sublett J, Kaushal D, Hinojosa E, Barton T, Tuomanen EI, et al. Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. Infect Immun 2006;74:4766–77.

[35] Silva NA, McCluskey J, Jefferies JM, Hinds J, Smith A, Clarke SC, et al. Genomic diversity between strains of the same serotype and multilocus sequence type among pneumococcal clinical isolates. Infect Immun 2006;74:3513–18.

[36] Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2006;2:e31.

[37] Calix JJ, Nahm MH. A new pneumococcal serotype, 11E, has a variably inactivated *wcjE* gene. J Infect Dis 2010;202:29–38.

[38] Jin P, Kong F, Xiao M, Oftadeh S, Zhou F, Liu C, et al. First report of putative *Streptococcus pneumoniae* serotype 6D among nasopharyngeal isolates from Fijian children. J Infect Dis 2009;200:1375–80.

[39] Park IH, Pritchard DG, Cartee R, Brandao A, Brandileone MC, Nahm MH. Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. J Clin Microbiol 2007;45:1225–33.

[40] Calix JJ, Porambo RJ, Brady AM, Larson TR, Yother J, Abeygunwardana C, et al. Biochemical, genetic, and serological characterization of two capsule subtypes among *Streptococcus pneumoniae* Serotype 20 strains: discovery of a new pneumococcal serotype. J Biol Chem 2012;287:27885–94.

[41] Aanensen DM, Mavroidi A, Bentley SD, Reeves PR, Spratt BG. Predicted functions and linkage specificities of the products of the *Streptococcus pneumoniae* capsular biosynthetic loci. J Bacteriol 2007;189:7856–76.
[42] Jonsson S, Musher DM, Chapman A, Goree A, Lawrence EC. Phagocytosis and killing of common bacterial pathogens of the lung by human alveolar macrophages. J Infect Dis 1985;152:4–13.
[43] Musher DM. Infections caused by *Streptococcus pneumoniae*: clinical spectrum, pathogenesis, immunity, and treatment. Clin Infect Dis 1992;14:801–9.
[44] Nelson AL, Roche AM, Gould JM, Chim K, Ratner AJ, Weiser JN. Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. Infect Immun 2007;75:83–90.
[45] van der Poll T, Opal SM. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. Lancet 2009;374:1543–56.
[46] Sa-Leao R, Simoes AS, Nunes S, Sousa NG, Frazao N, de LH. Identification, prevalence and population structure of non-typable *Streptococcus pneumoniae* in carriage samples isolated from preschoolers attending day-care centres. Microbiology 2006;152:367–76.
[47] Martin M, Turco JH, Zegans ME, Facklam RR, Sodha S, Elliott JA, et al. An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*. N Engl J Med 2003;348:1112–21.
[48] Scott JR, Hinds J, Gould KA, Millar EV, Reid R, Santosham M, et al. Nontypeable pneumococcal isolates among navajo and white mountain apache communities: are these really a cause of invasive disease? J Infect Dis 2012;206:73–80.
[49] Xu Q, Kaur R, Casey JR, Sabharwal V, Pelton S, Pichichero ME. Nontypeable *Streptococcus pneumoniae* as an otopathogen. Diagn Microbiol Infect Dis 2011;69:200–4.
[50] Hathaway LJ, Stutzmann MP, Battig P, Aebi S, Muhlemann K. A homologue of *aliB* is found in the capsule region of nonencapsulated *Streptococcus pneumoniae*. J Bacteriol 2004;186:3721–9.
[51] Park IH, Kim KH, Andrade AL, Briles DE, McDaniel LS, Nahm MH. Nontypeable pneumococci can be divided into multiple *cps* types, including one type expressing the novel gene *pspK*. MBio 2012;3.
[52] Keller LE, Jones CV, Thornton JA, Sanders ME, Swiatlo E, Nahm MH, et al. PspK of *Streptococcus pneumoniae* increases adherence to epithelial cells and enhances nasopharyngeal colonization. Infect Immun 2013;81:173–81.
[53] Salter SJ, Hinds J, Gould KA, Lambertsen L, Hanage WP, Antonio M, et al. Variation at the capsule locus, *cps*, of mistyped and non-typable *Streptococcus pneumoniae* isolates. Microbiology 2012;158:1560–9.
[54] Bergmann S, Hammerschmidt S. Versatility of pneumococcal surface proteins. Microbiology 2006;152:295–303.
[55] Schneewind O, Fowler A, Faull KF. Structure of the cell wall anchor of surface proteins in *Staphylococcus aureus*. Science 1995;268:103–6.
[56] Tong HH, Blue LE, James MA, DeMaria TF. Evaluation of the virulence of a *Streptococcus pneumoniae* neuraminidase-deficient mutant in nasopharyngeal colonization and development of otitis media in the chinchilla model. Infect Immun 2000;68:921–4.
[57] Trappetti C, Kadioglu A, Carter M, Hayre J, Iannelli F, Pozzi G, et al. Sialic acid: a preventable signal for pneumococcal biofilm formation, colonization, and invasion of the host. J Infect Dis 2009;199:1497–505.
[58] King SJ, Whatmore AM, Dowson CG. NanA, a neuraminidase from *Streptococcus pneumoniae*, shows high levels of sequence diversity, at least in part through recombination with *Streptococcus oralis*. J Bacteriol 2005;187:5376–86.
[59] Pettigrew MM, Fennie KP, York MP, Daniels J, Ghaffar F. Variation in the presence of neuraminidase genes among *Streptococcus pneumoniae* isolates with identical sequence types. Infect Immun 2006;74:3360–5.
[60] Smith A, Johnston C, Inverarity D, Slack M, Paterson GK, Diggle M, et al. Investigating the role of pneumococcal neuraminidase A activity in isolates from pneumococcal haemolytic uraemic syndrome. J Med Microbiol 2013;62:1735–42.
[61] Camara M, Mitchell TJ, Andrew PW, Boulnois GJ. *Streptococcus pneumoniae* produces at least two distinct enzymes with neuraminidase activity: cloning and expression of a second neuraminidase gene in *Escherichia coli*. Infect Immun 1991;59:2856–8.
[62] Banerjee A, van Sorge NM, Sheen TR, Uchiyama S, Mitchell TJ, Doran KS. Activation of brain endothelium by pneumococcal neuraminidase NanA promotes bacterial internalization. Cell Microbiol 2010;12:1576–88.
[63] Uchiyama S, Carlin AF, Khosravi A, Weiman S, Banerjee A, Quach D, et al. The surface-anchored NanA protein promotes pneumococcal brain endothelial cell invasion. J Exp Med 2009;206:1845–52.
[64] Meyer K, Chaffee E, Hobby GL, Dawson MH. Hyaluronidases of bacterial and animal origin. J Exp Med 1941;73:309–26.
[65] Irwin CR, Schor SL, Ferguson MW. Effects of cytokines on gingival fibroblasts in vitro are modulated by the extracellular matrix. J Periodontal Res 1994;29:309–17.
[66] McKee CM, Penno MB, Cowman M, Burdick MD, Strieter RM, Bao C, et al. Hyaluronan (HA) fragments induce chemokine gene expression in alveolar macrophages. The role of HA size and CD44. J Clin Invest 1996;98:2403–13.

B. GENETICS AND FUNCTIONAL GENOMICS OF *STREPTOCOCCUS PNEUMONIAE*

[67] Bethe G, Nau R, Wellmer A, Hakenbeck R, Reinert RR, Heinz H-P, et al. The cell wall-associated serine protease PrtA: a highly conserved virulence factor of *Streptococcus pneumoniae*. FEMS Microbiol Lett 2001;27:99–103.

[68] Mirza S, Wilson L, Benjamin Jr. WH, Novak J, Barnes S, Hollingshead SK, et al. Serine protease PrtA from *Streptococcus pneumoniae* plays a role in the killing of S. pneumoniae by apolactoferrin. Infect Immun 2011;79:2440–50.

[69] Hendriksen WT, Bootsma HJ, van DA, Estevao S, Kuipers OP, de GR, et al. Strain-specific impact of PsaR of *Streptococcus pneumoniae* on global gene expression and virulence. Microbiology 2009;155:1569–79.

[70] Kloosterman TG, Witwicki RM, van der Kooi-Pol MM, Bijlsma JJ, Kuipers OP. Opposite effects of Mn^{2+} and Zn^{2+} on PsaR-mediated expression of the virulence genes *pcpA*, *prtA*, and *psaBCA* of *Streptococcus pneumoniae*. J Bacteriol 2008;190:5382–93.

[71] Barocchi MA, Ries J, Zogaj X, Hemsley C, Albiger B, Kanth A, et al. A pneumococcal pilus influences virulence and host inflammatory responses. Proc Natl Acad Sci U S A 2006;103:2857–62.

[72] Hava DL, Hemsley CJ, Camilli A. Transcriptional regulation in the *Streptococcus pneumoniae rlrA* pathogenicity islet by RlrA. J Bacteriol 2003;185:413–21.

[73] Paterson GK, Mitchell TJ. The role of *Streptococcus pneumoniae* sortase A in colonisation and pathogenesis. Microbes Infect 2006;8:145–53.

[74] Bagnoli F, Moschioni M, Donati C, Dimitrovska V, Ferlenghi I, Facciotti C, et al. A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. J Bacteriol 2008;190:5480–92.

[75] Hoskins J, Alborn Jr WE, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, et al. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 2001;183:5709–17.

[76] Frolet C, Beniazza M, Roux L, Gallet B, Noirclerc-Savoye M, Vernet T, et al. New adhesin functions of surface-exposed pneumococcal proteins. BMC Microbiol 2010;10:190.

[77] Brown JS, Ogunniyi AD, Woodrow MC, Holden DW, Paton JC. Immunization with components of two iron uptake ABC transporters protects mice against systemic *Streptococcus pneumoniae* infection. Infect Immun 2001;69:6702–6.

[78] Brown JS, Gilliland SM, Ruiz-Albert J, Holden DW. Characterization of *pit*, a *Streptococcus pneumoniae* iron uptake ABC transporter. Infect Immun 2002;70:4389–98.

[79] Marra A, Lawson S, Asundi JS, Brigham D, Hromockyj AE. In vivo characterization of the psa genes from *Streptococcus pneumoniae* in multiple models of infection. Microbiology 2002;148:1483–91.

[80] Dintilhac A, Alloing G, Granadel C, Claverys J-P. Competence and virulence of *Streptococcus pneumoniae*: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. Mol Microbiol 1997;25:727–39.

[81] Loisel E, Jacquamet L, Serre L, Bauvois C, Ferrer JL, Vernet T, et al. AdcAII, a new pneumococcal Zn-binding protein homologous with ABC transporters: biochemical and structural analysis. J Mol Biol 2008;381:594–606.

[82] Bayle L, Chimalapati S, Schoehn G, Brown J, Vernet T, Durmort C. Zinc uptake by *Streptococcus pneumoniae* depends on both AdcA and AdcAII and is essential for normal bacterial morphology and virulence. Mol Microbiol 2011;82:904–16.

[83] Sutcliffe IC, Harrington DJ. Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. Microbiology 2002;148:2065–77.

[84] Chimalapati S, Cohen JM, Camberlein E, MacDonald N, Durmort C, Vernet T, et al. Effects of deletion of the *Streptococcus pneumoniae* lipoprotein diacylglyceryl transferase gene lgt on ABC transporter function and on growth in vivo. PLoS One 2012;7:e41393.

[85] Benton KA, Paton JC, Briles DE. Differences in virulence for mice among *Streptococcus pneumoniae* strains of capsular types 2, 3, 4, 5, and 6 are not attributable to differences in pneumolysin production. Infect Immun 1997;65:1237–44.

[86] Kirkham LA, Jefferies JM, Kerr AR, Jing Y, Clarke SC, Smith A, et al. Identification of invasive serotype 1 pneumococcal isolates that express nonhemolytic pneumolysin. J Clin Microbiol 2006;44:151–9.

[87] Jefferies JM, Johnston CH, Kirkham LA, Cowan GJ, Ross KS, Smith A, et al. Presence of nonhemolytic pneumolysin in serotypes of *Streptococcus pneumoniae* associated with disease outbreaks. J Infect Dis 2007;196:936–44.

[88] Lock RA, Zhang QY, Berry AM, Paton JC. Sequence variation in the *Streptococcus pneumoniae* pneumolysin gene affecting haemolytic activity and electrophoretic mobility of the toxin. Microb Pathog 1996;21:71–83.

[89] Witzenrath M, Pache F, Lorenz D, Koppe U, Gutbier B, Tabeling C, et al. The NLRP3 inflammasome is differentially activated by pneumolysin variants and contributes to host defense in pneumococcal pneumonia. J Immunol 2011;187:434–40.

[90] Kanclerski K, Mollby R. Production and purification of *Streptococcus pneumoniae* hemolysin (pneumolysin). J Clin Microbiol 1987;25:222–5.

B. GENETICS AND FUNCTIONAL GENOMICS OF *STREPTOCOCCUS PNEUMONIAE*

[91] Garnier F, Janapatla RP, Charpentier E, Masson G, Grelaud C, Stach JF, et al. Insertion sequence 1515 in the *ply* gene of a type 1 clinical isolate of *Streptococcus pneumoniae* abolishes pneumolysin expression. J Clin Microbiol 2007;45:2296–7.
[92] Berry AM, Lock RA, Hansman D, Paton JC. Contribution of autolysin to virulence of *Streptococcus pneumoniae*. Infect Immun 1989;57:2324–30.
[93] Balachandran P, Hollingshead SK, Paton JC, Briles DE. The autolytic enzyme LytA of *Streptococcus pneumoniae* is not responsible for releasing pneumolysin. J Bacteriol 2001;183:3116.
[94] Price KE, Camilli A. Pneumolysin localizes to the cell wall of *Streptococcus pneumoniae*. J Bacteriol 2009;191:3651–60.
[95] Price KE, Greene NG, Camilli A. Export requirements of pneumolysin in *Streptococcus pneumoniae*. J Bacteriol 2012;194:3651–60.
[96] Friedland IR, Paris MM, Hickey S, Shelton S, Olsen K, Paton JC, et al. The limited role of pneumolysin in the pathogenesis of pneumococcal meningitis. J Infect Dis 1995;172:805–9.
[97] Wellmer A, Zysk G, Gerber J, Kunst T, Von MM, Bunkowski S, et al. Decreased virulence of a pneumolysin-deficient strain of *Streptococcus pneumoniae* in murine meningitis. Infect Immun 2002;70:6504–8.
[98] Winter AJ, Comis SD, Osborne MP, Tarlow MJ, Stephen J, Andrew PW, et al. A role for pneumolysin but not neuraminidase in the hearing loss and cochlear damage induced by experimental pneumococcal meningitis in guinea pigs. Infect Immun 1997;65:4411–18.
[99] Benton KA, Everson MP, Briles DE. A pneumolysin-negative mutant of *Streptococcus pneumoniae* causes chronic bacteremia rather than acute sepsis in mice. Infect Immun 1995;63:448–55.
[100] Canvin JR, Marvin AP, Sivakumaran M, Paton JC, Boulnois GJ, Andrew PW, et al. The role of pneumolysin and autolysin in the pathology of pneumonia and septicemia in mice infected with a type 2 pneumococcus. J Infect Dis 1995;172:119–23.
[101] Zysk G, Schneider-Wald B, Hwang JH, Bejo L, Kim KS, Mitchell T, et al. Pneumolysin is the main inducer of cytotoxicity to brain microvascular endothelial cells caused by *Streptococcus pneumoniae*. Infect Immun 2000;69:845–52.
[102] Lange R, Wagner C, de Saizieu A, Flint N, Molnos J, Stieger M, et al. Domain organization and molecular characterization of 13 two-component systems identified by genome sequencing of *Streptococcus pneumoniae*. Gene 1999;237:223–34.
[103] Throup JP, Koretke KK, Bryant AP, Ingraham KA, Chalker AF, Ge Y, et al. A genomic analysis of two-component signal transduction in *Streptococcus pneumoniae*. Mol Microbiol 2000;35:566–76.
[104] Blue CE, Mitchell TJ. Contribution of a response regulator to the virulence of *Streptococcus pneumoniae* is strain dependent. Infect Immun 2003;71:4405–13.
[105] Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, Feltwell T, et al. Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. PLoS Genet 2013;9:e1003868.
[106] Overturf GD. Infections and immunizations of children with sickle cell disease. Adv Pediatr Infect Dis 1999;14:191–218.
[107] Darghouth D, Koehl B, Madalinski G, Heilier JF, Bovee P, Xu Y, et al. Pathophysiology of sickle cell disease is mirrored by the red blood cell metabolome. Blood 2011;117:e57–66.
[108] Prasad AS, Lei KY, Moghissi KS, Stryker JC, Oberleas D. Effect of oral contraceptives on nutrients. III. Vitamins B6, B12, and folic acid. Am J Obstet Gynecol 1976;125:1063–9.
[109] Carter R, Wolf J, van OT, Muller M, Obert C, Burnham C, et al. Genomic analyses of pneumococci from children with sickle cell disease expose host-specific bacterial adaptations and deficits in current interventions. Cell Host Microbe 2014;15:587–99.
[110] Arbique JC, Poyart C, Trieu-Cuot P, Quesne G, Carvalho MGS, Steigerwalt AG, et al. Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. J Clin Microbiol 2004;42:4686–96.
[111] Shahinas D, Thornton CS, Tamber GS, Arya G, Wong A, Jamieson FB, et al. Comparative genomic analyses of provide insight into virulence and commensalism dynamics. PLoS One 2013;8:e65670.
[112] Rolo D, Simoes S, Domenech A, Fenoll A, Linares J, de LH, et al. Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain. PLoS One 2013;8:e57047.
[113] Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG. Assigning strains to bacterial species via the internet. BMC Biol 2009;7:3.
[114] Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, et al. Evolution of *Streptococcus pneumoniae* and its close commensal relatives. PLoS One 2008;3:e2683.
[115] Chi F, Nolte O, Bergmann C, Ip M, Hakenbeck R. Crossing the barrier: evolution and spread of a major class of mosaic *pbp2x* in *S. pneumoniae*, *S. mitis* and *S. oralis*. Int J Med Microbiol 2007;297:503–12.

B. GENETICS AND FUNCTIONAL GENOMICS OF *STREPTOCOCCUS PNEUMONIAE*

[116] Morrison DA, Lee MS. Regulation of competence for genetic transformation in *Streptococcus pneumoniae*: a link between quorum sensing and DNA processing genes. Res Microbiol 2000;151:445–51.

[117] Kilian M, Riley DR, Jensen A, Bruggemann H, Tettelin H. Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. MBio 2014;5:e01490–14.

[118] Sauerbier J, Maurer P, Rieger M, Hakenbeck R. *Streptococcus pneumoniae* R6 interspecies transformation: genetic analysis of penicillin resistance determinants and genome-wide recombination events. Mol Microbiol 2012;86:692–706.

[119] Denapaite D, Brückner R, Nuhn M, Reichmann P, Henrich B, Maurer P, et al. The genome of *Streptococcus mitis* B6—what is a commensal? PLoS One 2010;5:e9426.

[120] Reichmann P, Nuhn M, Denapaite D, Brückner R, Henrich B, Maurer P, et al. Genome of *Streptococcus oralis* strain Uo5. J Bacteriol 2011;193:2888–9.

[121] Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell B, Parkhill J. ACT: the artemis comparison tool. Bioinformatics 2005;21:3422–3.

[122] Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 2000;1:RESEARCH0011.

[123] Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 2004;5:150–63.

[124] Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russel JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 1998;95:3140–5.

[125] Shahinas D, Tamber GS, Arya G, Wong A, Lau R, Jamieson F, et al. Whole-genome sequence of *Streptococcus pseudopneumoniae* isolate IS7493. J Bacteriol 2011;193:6102–3.

[126] Hakenbeck R, Balmelle N, Weber B, Gardes C, Keck W, de Saizieu A. Mosaic genes and mosaic chromosomes: intra- and interspecies variation of *Streptococcus pneumoniae*. Infect Immun 2001;69:2477–86.

[127] Schähle Y. Genomische diversität und evolution von virulenzdeterminanten in *Streptococcus* spp. 2010.

[128] Chi F, Leider M, Leendertz F, Bergmann C, Boesch C, Schenk S, et al. New *Streptococcus pneumoniae* clones in deceased wild chimpanzees. J Bacteriol 2007;189:6085–8.

[129] Denapaite D, Hakenbeck R. A new variant of the capsule 3 cluster occurs in *Streptococcus pneumoniae* from deceased wild chimpanzees. PLoS One 2011;e25119. Epub 2011 Sep 28.

[130] Whatmore AM, King SJ, Doherty NC, Sturgeon D, Chanter N, Dowson CG. Molecular characterization of equine isolates of *Streptococcus pneumoniae*: natural disruption of genes encoding the virulence factors pneumolysin and autolysin. Infect Immun 1999;67:2776–82.

[131] Sibold C, Henrichsen J, König A, Martin C, Chalkley L, Hakenbeck R. Mosaic *pbpX* genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from *pbpX* genes of a penicillin-sensitive *Streptococcus oralis*. Mol Microbiol 1994;12:1013–23.

[132] Dowson CG, Coffey TJ, Kell C, Whiley RA. Evolution of penicillin resistance in *Streptococcus pneumoniae*; the role of *Streptococcus mitis* in the formation of a low affinity PBP2B in *S. pneumoniae*. Mol Microbiol 1993;9:635–43.

[133] Denapaite D, Bruckner R, Hakenbeck R, Vollmer W. Biosynthesis of teichoic acids in *Streptococcus pneumoniae* and closely related species: lessons from genomes. Microb Drug Resist 2012;18:344–58.

[134] Hakenbeck R, Madhour A, Denapaite D, Brückner R. Versatility of choline metabolism and choline binding proteins in *Streptococcus pneumoniae* and commensal streptococci. *FEMS*. Microbiol Rev 2009;33:572–86.

[135] Madhour A, Maurer P, Hakenbeck R. Cell surface proteins in *S. pneumoniae*, *S. mitis* and *S. oralis*. Iran J Microbiol 2011;3:58–67.

[136] Takahashi Y, Konishi K, Cisar JO, Yoshikawa M. Identification and characterization of hsa, the gene encoding the sialic acid-binding adhesin of *Streptococcus gordonii* DL1. Infect Immun 2002;70:1209–18.

[137] Aguiar SI, Serrano I, Pinto FR, Melo-Cristino J, Ramirez M. The presence of the pilus locus is a clonal property among pneumococcal invasive isolates. BMC Microbiol 2008;8:41.

[138] Chandra RR, Dwivedi VP, Chatterjee S, Raghava Prasad DV, Das G. Early secretory antigenic target-6 of *Mycobacterium tuberculosis*: enigmatic factor in pathogen-host interactions. Microbes Infect 2012;14:1220–6.

[139] Brückner R, Nuhn M, Reichmann P, Weber B, Hakenbeck R. Mosaic genes and mosaic chromosomes—genomic variation in *Streptococcus pneumoniae*. Int J Med Microbiol 2004;294:157–68.

[140] Whatmore AM, Efstratiou A, Pickerill AP, Broughton K, Woodward G, Sturgeon D, et al. Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. Infect Immun 2000;68:1374–82.

[141] Kearns AM, Wheeler J, Freeman R, Seiders PR, Perry J, Whatmore AM, et al. Pneumolysin detection identifies atypical isolates of *Streptococcus pneumoniae*. J Clin Microbiol 2000;38:1309–10.

3 Unpublished material

This chapter describes additional details of the publications described in chapter 2 and further unpublished work.

3.1 Analysis of *Streptococcus pneumoniae* clone ST10523

The annotated genome sequences of the *S. pneumoniae* isolates D122, D141 and D219, all representatives of the new clone ST10523, were generated and analysed as described in the publication (see chapter 2.1). The final genome sequences were based on 166.138 – 209.299 sequence reads with 30.782.842 – 41.092.696 nucleotides (nt) obtained with 454 sequencing technology (supplementary table S1). 25.631.541 (D122), 34.737.088 (D141) and 28.908.339 nt (D219) of reads were assembled using paired-end information into 2.066.903 (D122), 2.075.725 (D141) and 2.092.317 nt (D219) in 340 (D122), 181 (D141) and 413 (D219) contigs. Genome comparisons and SNP analyses were performed using a program especially developed for this purpose (see chapter 3.6). The main questions here were whether there are differences between the two strains isolated from one patient over three years apart (*S. pneumoniae* D121 and D141), and how these two strains differed from strain *S. pneumoniae* D219 isolated independently at a different time and place. Furthermore, to reveal possible clone specific genes, the deduced proteins were compared to those of other *S. pneumoniae* clones. Since the 23F-capsule cluster of the ST10523 isolates, a major virulence factor, was excluded from first analysis step due to variability of this locus among *S. pneumon*iae, it was compared in detail to the capsule of the serotype 23F reference clone *S. pneumoniae* ATCC700669 (Acc.No. NC_011900; afterwards referred to as 23F). This chapter shows details of this work which had not been included in the publication.

3.1.1 Genome comparison

3.1.1.1 Regions of divergent sequences

Pairwise comparisons between the *S. pneumoniae* D122, D141 and D219 genomes revealed a varying number of regions that differed between two genomes: 252 regions between D122 and D141, 303 regions between D141 and D219 and 415 regions between D122 and D219. Most of these regions were detected at contig edges that defined sequence gaps in one genome, where the gap is represented by a given number of N (N-stretch) (Figure 3.1) introduced automatically by the assembly program *Newbler* (*gsAssembler*) (Margulies, et al., 2005) version 2.6. Since the genome sequences of the ST10523 strains originated from paired-end-sequencing, assembled contigs could be combined to scaffolds, interspersed by gaps of unknown sequence. The number of N (representing the gap) is determined by the distance of the generated contigs as given by the read-pair information. Gaps between scaffolds, which have been arranged in regard to a reference sequence (*S. pneumoniae* R6 for D219 and D219 for D122 and D141), are represented by one hundred N according to the current NCBI guidelines (NCBI). These and inter-contig N-stretches within scaffolds contain no information and, consequently, were neglected in all further analyses. Regions which contain N-stretches were inspected manually and removed from the analysis. Furthermore, regions with duplicated sequences at contig edges were also removed from analysis. Using pairwise genome comparison of the three genomes, between two and five regions per comparison at a total of seven locations (Figure 3.2) were analysed in detail. It is important to realize that each genome was composed of a distinct number of contigs, and gaps occurred not necessarily at the same positions in the genomes. The pairwise comparisons were combined into a comparison of all three genomes, where the number of regions to be investigated increased accordingly. The comparison of D122 and D141 revealed two differing regions. Only one region remained in the overall comparison with D219, since sequence information concerning the second gap was too limited in D219. Furthermore, D122 and D141 differed from D219 in four regions. A fifth region again could not be used in the overall comparison due to gaps. Thus, the genome sequence common to all three genomes is smaller than the estimated genome size of each ST10523 strain.

Two of these regions, where both, D122 and D141, differed from D219, included the phage relict and the prophage present only in D219 as described in the publication.

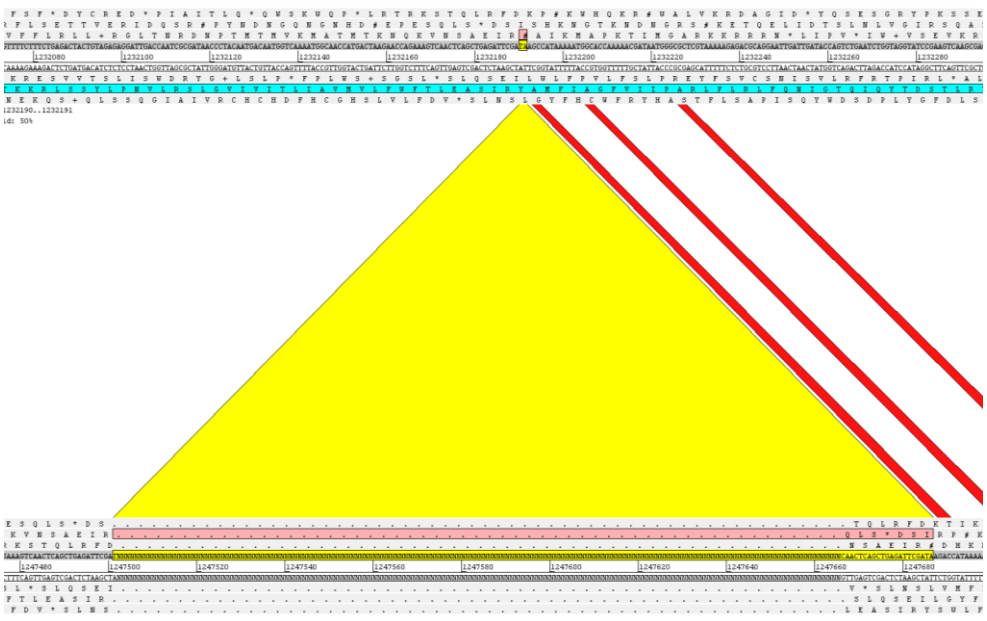


Figure 3.1: Representation of a discontinuous region in one genome in ACT genome comparison

The visualization with the Artemis comparison tool (ACT) shows an example of the alignment between *S. pneumoniae* D141 (**top**) and D122 (**bottom**) genome sequences. The D141 sequence is contiguous, whereas this region is discontinuous in case of D122, i.e. is located on two contigs. The apparent gap between the two D122 contigs is filled by a stretch of N (representing unknown sequence) during the assembly with Newbler 2.6. In the ACT visualization, this region might be mistaken as an insertion in D122. Frequently, sequences flanking such gaps are ambiguous or duplicated, depending on the sequencing and assembly technology and the assembly procedure used. For resolution of this problem, see text for details.

A region of approximately 7.500 nt absent in the D219 genome, located between SPND219_00722 and SPND219_00724 and representing five genes in the genomes of D122 and D141, is also described in the publication. In this case, the sequence of D122 and D141 is distributed on several small contigs (Figure 3.3) and therefore contains N-stretches of approximately 1.700 N. In D219, 95 nt remained at one contig edge with no match to D122 but to D141.

Furthermore, D122 contains an exchange of 433 nt by 369 nt within the gene SPND122_00874 (encoding the specificity subunit S of type I restriction-modification system; the gene is incomplete at a gap in the D219 genome) compared to the other two genomes. This protein family is known to be highly variable, and as discussed by Manso *et al.* (Manso, et al., 2014) plays a role as an important regulator of pneumococcal virulence by enabling phenotype switching between opaque and transparent colony morphology (Weiser, et al., 1994). Although rearrangements within the encoding genes are mentioned in this publication, the second half of the affected gene appears to be entirely different in D141 compared to D122.

The exchanged region is most likely due to an assembly error, since the D122-sequence (433 nt) is present approximately 2.500 nt downstream in case of D141, whereas D122 contains a gap at the corresponding location (Figure 3.5).

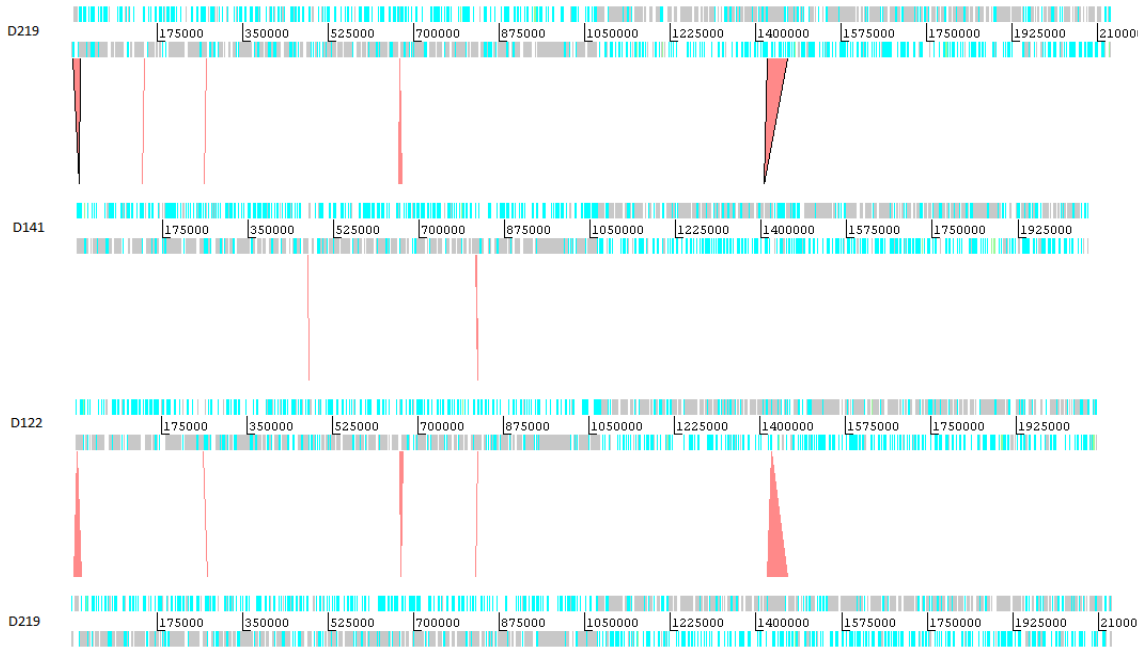


Figure 3.2: Unaligned regions in pairwise comparison of ST10523 genomes

The visualization with the Artemis comparison tool (ACT) shows an overview of the regions obtained from a pairwise alignment of the ST10523 genomes. Regions containing only or mainly N or sequence which is duplicated at contig edge were removed.

Sixty-three nt downstream of the gene SPND141_00511, the D141 sequence reveals a deletion of 829 nt, containing fragments of a gene encoding a type I restriction endonuclease (Figure 3.4). Since this is located 227 nt upstream of a gap, this region was not further analysed due to potentially low sequence quality near contig edges especially of sequences generated by the 454-sequencing technology.

The remaining two regions are located at BOX elements (see introduction, chapter 1.1.1) and due to assembly problems at repeats they were also not considered.

In summary, no differences in gene content between the three genomes were apparent except for the phage relict and the prophage present only in D219 as well as a gapped five-gene-cluster missing in the genome of D219 between SPND219_00722 and SPND29_00724. The fact that D219 differs more to each of the two strains D122 and D141 than D122 and D141 to each other is easily explainable by the fact that D122 and D141 were isolated from the same

patient and D219 was isolated at a different time and location from another host. Differences based on the presence of SNPs are described in chapter 3.1.2.

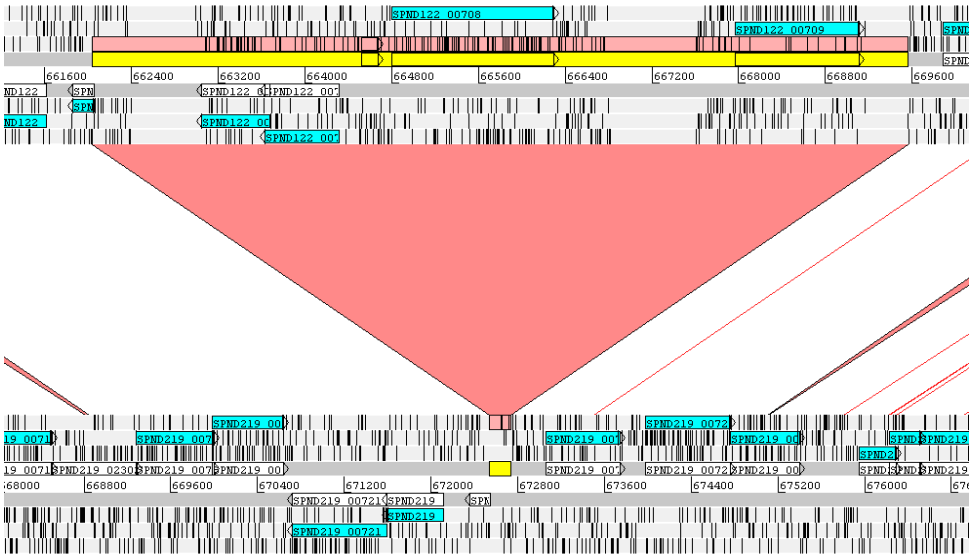


Figure 3.3: Genes absent in D219

The visualization with Artemis comparison tool (ACT) shows the alignment of Streptococcus pneumoniae D219 (**bottom**) and D122 (**top**) where five genes are present in D122. The region in D122 spans over several small contigs and contains approximately 23% of N bridging the contig gaps. The region looks similar in D141 with a different number of N. In D219, 95nt not matching D122 sequences remain at one contig edge (marked **yellow** in D219, selection contains a gap in D219).

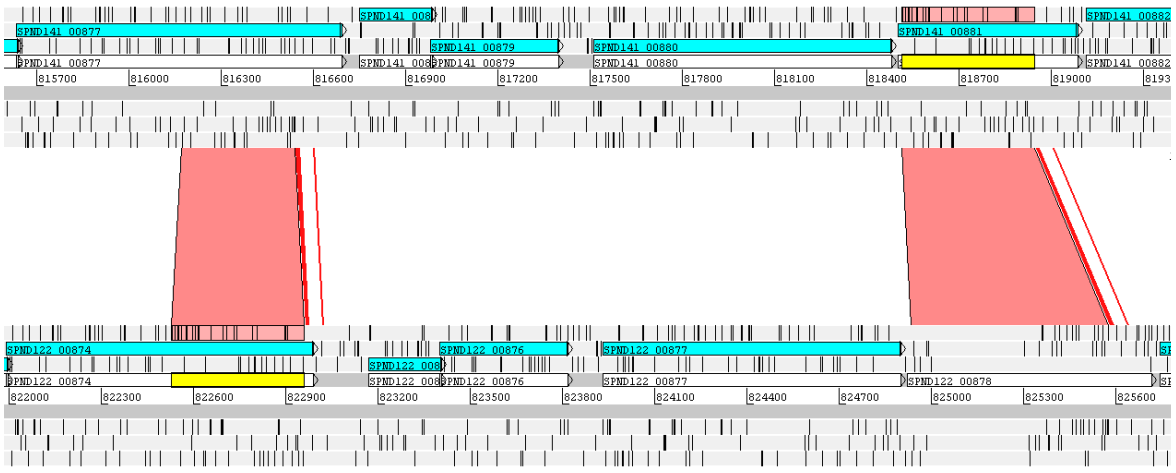


Figure 3.5: An apparent divergent region in SPND122_00874

Visualization with the Artemis comparison tool (ACT) shows the alignment of *Streptococcus pneumoniae* D122 (**bottom**) and D141 (**top**) at the location of the gene SPND122_00874/SPND141_00877 and downstream region. The altered region of SPND122_00874 (yellow) is present (also marked **yellow**) in D141 about 2.500 nt downstream of the exchanged region, where the sequence of D122 has a gap. The sequence in D122 is likely to be the result of a mis-assembly.

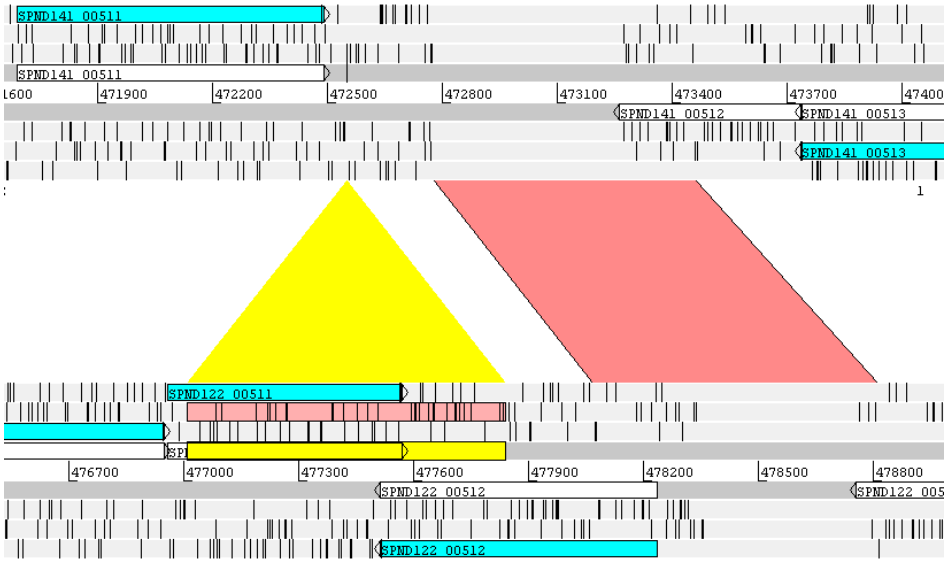


Figure 3.4: Apparent insertion in D122

The visualization with Artemis comparison tool (ACT) shows the alignment of *Streptococcus pneumoniae* D122 (**bottom**) and D141 (**top**) at the location of the gene SPND141_00511 and downstream region. 63 nt downstream of the gene SPND141_00511 829 nt are absent compared to D122 corresponding to a location 227 nt upstream of a sequence gap. This sequence contains fragments of a gene encoding a type I restriction modification system.

3.1.1.2 ST20523-specific genes

In order to see whether some gene products are specifically associated with the ST10523 clone, comparative analysis with other *S. pneumoniae* genomes, based on protein coding genes (CDS) common to all ST10523 genomes, were performed. RNA genes were excluded, since they are organized in several clusters with similar or nearly identical sequences, increasing the probability of assembly errors. CDS located in the serotype specific capsule cluster, where all three sequences are identical except for gaps, as well as in the two phage clusters, were analysed separately and were also excluded in the overall analysis, as well as small non-coding RNA genes. Furthermore, 207 – 233 CDS containing differences due to homopolymer stretches and partial CDS (incomplete at gap) in any of the three genomes were omitted as well as insertion sequences, transposases and mobile or repetitive elements (see introduction for further information). Each of the ST10253 genomes contained between 2.262 and 2.359 genes. After subtracting the genes described above, there remained a total of 1.591 – 1.620 genes per genome, which were included in the analysis. This means, more than 22% of all CDS and about 30% of all genes (31% of D219 due to phage clusters) are not used for analysis.

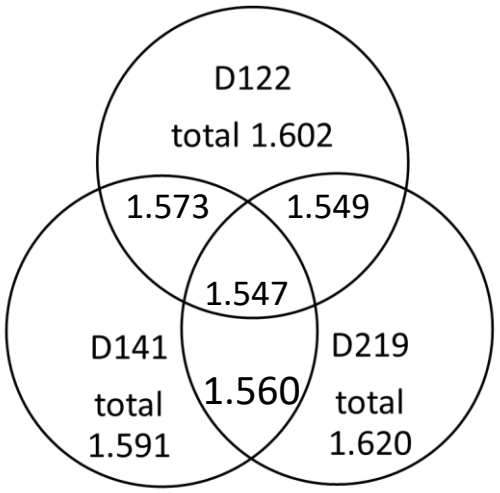


Figure 3.6: Comparison of protein coding genes of S. pneumoniae D122, D141 and D219.

The values represent proteins with at least 60% coverage and at least 70% identity. The analysis is based on all deduced proteins complete in all three isolates and not located within the capsule or any of the two phage clusters. Furthermore, proteins encoded by genes containing putative homopolymer differences or encoding transposases, insertion sequences, mobile or repetitive elements (e.g. BOX) were omitted. The differences between the three genomes arise mainly from SNVs in genes resulting in frameshifts.

Deduced protein sequences were used rather than DNA sequences. A coverage of ≥ 60% and an identity of ≥ 70% was used to define the presence of a protein. These values were chosen according to Denapaite *et al.* (Denapaite, et al., 2010) to allow a certain variability of the proteins which is important especially in comparisons between genomes of unrelated clones as described below. According to these values, a total of 1.547 representing 95 – 97% of the

1.591 – 1.620 genes were used for comparative analysis of other *S. pneumoniae* genomes, since they are common to all three ST10523 genomes. Figure 3.6 summarizes the result obtained by the pairwise comparison between all three ST10523 genomes. Between 28 and 73 proteins consist of fragments due to frameshifts in their genes and thus appear to be absent in some genomes although the DNA sequence was present. Based on these numbers, a pan-genome of these three genomes was estimated with 1.678 proteins/protein coding genes. The 1.547 proteins common to all genomes, represent 92 % of the pan-genome (95 – 97% of individual genomes). 5 % (90 proteins) of the pan-genome seems "unique" to any of the three genomes and 2% (41 proteins) are shared by two genomes. It should be kept in mind, that most of the "unique" proteins are fragments caused by frameshift or premature stop. Comparison of their encoding nucleotide sequence reveals their presence in all compared isolates.

Only one protein differs in *S. pneumoniae* D219 (SPND219_00891) and D141 (SPND141_00877) compared to D122 (SPND122_00874) in the gene encoding a specificity subunit S of a type I restriction-modification system. Besides some SNPs and indels described below, a sequence exchange already mentioned is apparent.

In the next step, the 1.547 proteins were compared to those of the laboratory strain

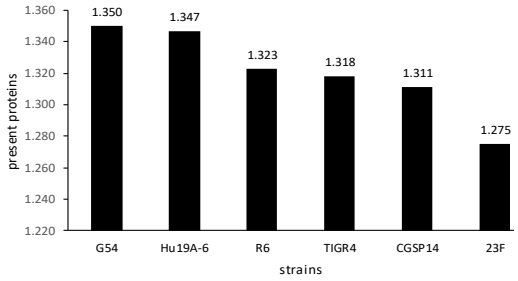


Figure 3.7: D219-proteins present in other S. pneumoniae strains

Based on a coverage of ≥60% and an identity of ≥70%, six *S. pneumoniae* strains (R6, TIGR4, ATCC 700669, Hu^{19A}-6, G54 and CGSP14) contain between 82 - 87% of the 1.547 D219 deduced proteins, corresponding to 1.275 proteins of 23F and up to 1.350 proteins of G54.

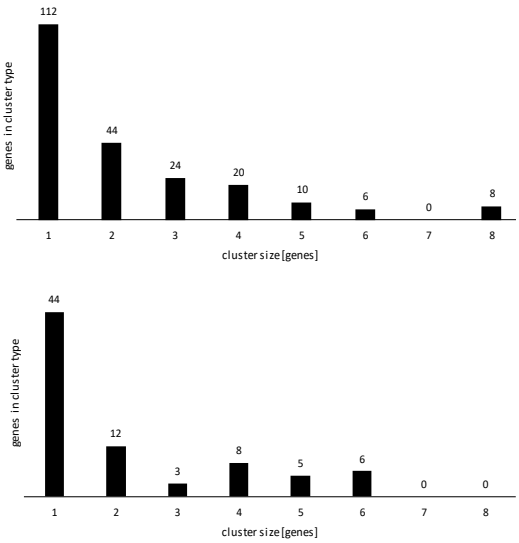


Figure 3.8: Clustering of protein coding genes in D219, absent in other strains

Genes encoding D219 proteins which were below a coverage of 60% and an identity of 70% in other *S. pneumoniae* strains are frequently organized in clusters.

Upper chart: 224 of the deduced D219 proteins are not present in *S. pneumoniae* R6. 50% of them are organized in clusters between two to eight genes.

Lower chart: 78 of the deduced D219 proteins are not present in any of the six *S. pneumoniae* genomes (R6, TIGR4, ATCC 700669, Hu^{19A}-6, G54 or CGSP14). Thirty-four (43%) of them are organized in clusters of two to six genes.

S. pneumoniae R6 which was used as reference genome. For this purpose, deduced proteins with a coverage of at least 60% and an identity of at least 70% were retrieved by a combination of *BLASTP* (Altschul, et al., 1990) and *Clustalw2* (Thompson, et al., 1994). It should be kept in mind that the annotation can differ between *S. pneumoniae* genomes, resulting in absence of CDS in some cases, although the DNA is present. Also, the product description as part of the annotation might differ while the same sequence is annotated. 1.323 (86%) of the 1.547 ST10523 proteins are present in *S. pneumoniae* R6. In case of the D219 genome, there remained 224 proteins whose genes were spread throughout the genome with 112 (50%) of them being organized in 39 clusters composed of between two to eight genes (Figure 3.8, upper part).

This type of comparison was extended to the genomes of the *S. pneumoniae* strains TIGR4 (accession number NC_003028), Hu^{19A}-6 (NC_010380), CGSP14 (NC_01058), G54 (NC_011072) and ATCC 700669 (NC_011900) which is referred to as 23F in the current work. These strains were chosen to provide a diverse set of strains of distinct genotype, isolated at various locations. A total of 1.257 – 1.350 (82 – 87%) of the 1.547 ST10523 proteins are present in the six genomes (Figure 3.7), 1.146 genes are present in all genomes. Only 78 *S. pneumoniae* D219 proteins could not be found in any of these strains. Similar to the corresponding genes not found in *S. pneumoniae* R6, 34 (44%) of the genes encoding the *S. pneumoniae* D219-specific proteins show a clustering into eleven groups of two to six genes (Figure 3.8, lower part).

A subsequent *BLAST* search of the 78 proteins against the NCBI database revealed hits with a similarity of 99 - 100% for every single one of these proteins in the thousands complete or incomplete *S. pneumoniae* genomes. Thus, none of the 1.547 proteins is specific to ST10523.

3.1.2 SNPs and indels in ST10523

The detailed retrieval of single nucleotide polymorphisms (SNPs) and single nucleotide deletions and insertions (indels), was performed by the software described in chapter 3.6 between the ST10523 genomes. As mentioned above, not considered in this analysis was a subset of genes such as IS elements, repetitive elements and RNA coding genes as well as differences in homopolymer stretches, incomplete genes and genes with differences closely (≤350 nt) located to contig edges.

Concerning the comparison of only *S. pneumoniae* D122 and D141, 63 genes in D122 and 56 genes in D141 contained 46 SNPs and 39 indels (see supplementary table S2). Six genes contained only silent SNPs. 38 genes in D122 and 31 genes in D141 were affected by frameshifts as a result of indels (in addition to potential SNPs in these genes) and 19 genes contained amino acid changing SNPs and no frameshift. Two of these genes contained SNPs, which either result in a stop codon (leading to a premature stop) or where a stop codon is affected to extend the coding region. These genes encode the substrate-binding component MalE of the maltose/maltodextrin ABC transporter and a hypothetical protein. The gene SPND122_00874 contained two indels and three SNPs directly after an exchanged region already described in chapter 3.1.1.1. It should be kept in mind, that the exchanged region of this gene is likely to be the result of a mis-assembly. Likewise, the for members of the same clone high number of potential frameshifts distributed all over the genome – also considering the underlying sequencing technology – indicates, that subsequent verification of them seems necessary to confirm or falsify the presence of authentic frameshifts.

Concerning the comparison of all three ST10523 genomes, 159 genes in D122, 153 genes in D141 and 163 genes in D219 contain 163 SNPs and 72 indels (see supplementary table S3), where 19 genes contain only silent SNPs. Counting CDS containing SNVs for each strain represents the relational distance of the isolates (Table 3.1). D122 and D141 contain only few affected CDS each (D122 compared to D141 and D219, which are equal at the

Table 3.1: CDS of all ST10523 isolates affected by SNVs

Counting CDS, which contain differences between the three isolates, represents their relational distance. Where D122 and D141 each show with 7 – 33 CDS containing SNVs only a low number of differences to both other isolates, the difference of D219 to the other with two is quite high with 111 – 119 affected CDS. Only 9 – 7 CDS differ in all three isolates.

differing isolate	affected CDS		
	D122	D141	D219
D122	33	28	
D141	7		
D219	111		119
all	7		9

particular CDS and so on). 7 – 33 CDS are affected by SNVs, where D122 or D141 are equal to D219. As expected by the relational distance, D219 contains a much higher number of SNVs (111 – 119 CDS) compared to D122 and D141. Only 7 – 9 CDS contain SNVs in all three isolates.

Another two genes are mentioned in the publication since they affect important gene products. They were excluded in the overall analysis since they are located close to contig edges; however, manual inspection of the sequences in these two regions implied true changes rather than sequence errors. This concerns the disruption of the non-essential histidine kinase SPND122_00180/SPND219_00200 in D141 due to the insertion of a transposase fragment. Moreover, the hyaluronidase gene *hlyA* contains a gap of four nt in the ST10523 genomes as well as a deletion in the promoter region, resulting in a non-functional gene product.

In addition, the IgA1-protease gene in D141 contains two non-silent SNVs and one indel, resulting in two adjacent amino acid exchanges ($KF^{D122}{}_{65}QL^{D141}$) a frameshift and thus a longer IgA1-gene (5.892 nt; 1.998 nt in D122). IgA1-proteases cleave immunoglobulin A1 to evade host immune defence (Chi, et al., 2017). The D219 allele was excluded from the analysis of all three isolates because of incompleteness of one gene fragment. Due to the sequencing technology used here, SNVs of potential interest should be verified by direct sequencing.

In conclusion, only the unusual *hlyA* present in the ST10523 clone possibly contributes to the ability to persist within the host over a long time period. The absence of two phage related regions in D141 and D122 could add to this property.

3.1.3 The capsule cluster

The ST10253 clone expresses a 23F capsule (Reichmann, et al., 1995). Although the 23F capsule is not associated with a high prevalence to cause invasive diseases (Croucher, et al., 2009; Sjöström, et al., 2006), this cluster was investigated since a reduced capacity to synthesize a capsule might contribute to the long persistence in a host of D141 and D122. No differences could be found between the capsule clusters of *S. pneumoniae* D219, D122 and D141, excluding differences within a region of 350 nt at contig edges and homopolymer stretches. The ST10523 isolates expressed a variant of the 23F-capsule when compared to the *cps* cluster of *S. pneumoniae* ATCC 700669 (Acc. No. NC_011900; afterwards referred to as 23F), which belongs to the Spain $23F^{-1}$ clone (ST81) (Croucher, et al., 2009).. Some members of the clone Spain$23F^{-1}$ express a different serotype due to a capsule switch which results in evasion of vaccine treatment (Croucher, et al., 2011; Croucher, et al., 2009; Coffey, et al., 1998; Coffey, et al., 1998; Klugman, 2002), but this is not an issue of the ST10523 genomes. The high prevalence worldwide of Spain$23F^{-1}$ is most likely due to its multiple antibiotic and high-level penicillin resistance phenotype.

For further analysis, the region between the genes encoding *DexB* and *AliA* was used for comparison and flanking mobile elements and transposases were excluded. The comparison of *S. pneumoniae* D219 with the capsule of *S. pneumoniae* 23F using the tool described in chapter 3.6 revealed several differences (Table 3.2). Most differences (Figure 3.9) are silent SNPs and thus do not affect the function of the gene product. Interestingly, the *rlmB* gene contains 22 SNPs (SPN23F_03350/SPND219_00397, dTDP-glucose-4,6-dehydratase), but all of them are silent (see publication for details). Thirteen amino acid changes in six genes remained.

There are only few differences between the ST10523- and the ST81-capsule which affect the encoded amino acid sequences. A quantitative analysis of the cell wall polysaccharide will be required to assess an effect on protein function and thus on capsule synthesis (Table 3.2).

Table 3.2: Divergences in proteins of the 23F capsule cluster of ST10523 and S. pneumoniae 23F

The capsule cluster of ST10523, compared to the capsule cluster of *S. pneumoniae* ATCC700669 (23F), contains only few differences with effect on the encoded proteins; Differences at contig edges and homopolymer stretches were not considered as well as intergenic differences. Only six genes contain a total of thirteen amino acid changing SNPs compared to 23F, while the capsule clusters of the ST10523 among each other are identical.

23F locus_tag	gene	product	D219 locus_tag	difference
SPN23F_03180	wzg	cps biosynthesis integral membrane regulatory protein Wzg	SPND219_00380	4 SNPs: silent, $I_{29}V$, $L_{80}V$, $E_{123}D$
SPN23F_03220	wchA	UDP-phosphate glucose phosphotransferase	SPND219_00384	2 SNPs: E_3G, $G_{202}S$
SPN23F_03290	wchX	glycerol phosphotransferase WchX	SPND219_00391	1 SNP: $G_{376}R$
SPN23F_03310	wchZ	nucleotidyl transferase WchZ	SPND219_00393	4 SNPs: silent, $I_{132}T$
SPN23F_03340	rmlC	dTDP-4-keto-6-deoxyglucose-3,5-epimerase RmlC	SPND219_00396	3 SNPs: silent, $C_{20}G$, $N_{182}H$
SPN23F_03360	rmlD	dTDP-4-dehydrorhamnose reductase RmlD	SPND219_00398	7 SNPs: silent, $A_{32}V$, $N_{38}D$, $E_{39}A$, $R_{87}K$

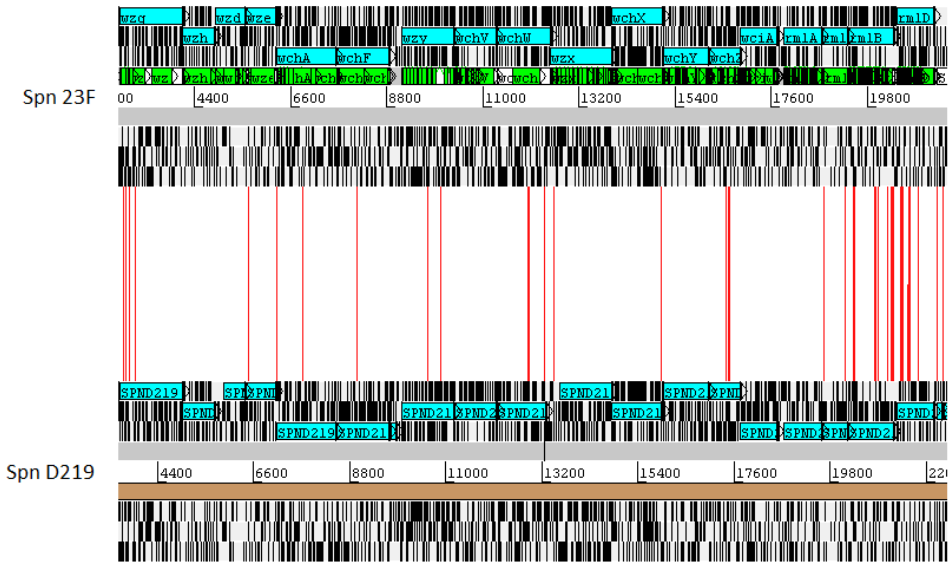


Figure 3.9: Comparison of the capsule cluster of ST10523 and S. pneumoniae ATCC700669

The visualization with Artemis comparison tool (ACT) shows the alignment of *Streptococcus pneumoniae* D219 (**bottom**) and ATCC700669 (23F) (**top**) at the location of the capsule cluster. Most of the differences represent silent SNPs. Only six genes contain SNPs which lead to thirteen amino acid changes, excluding differences in homopolymer stretches or near gap.

3.2 Analysis of *Streptococcus pneumoniae* clone ST226

The annotated genome sequences of the *S. pneumoniae* isolates Hu15 (penicillin sensitive) and Hu17 (penicillin resistant), both representatives of the ST226 clone, were generated and analysed as described in the publication (see chapter 2.2). Illumina sequencing technology resulted in 2.230.196 – 2.246.554 sequence reads with 336.759.596 – 339.229.654 nucleotides obtained with (supplementary table S1). 309.345.207 (Hu15) and 307.838.725 nt (Hu17) of reads were assembled using paired-end information into 2.136.165 (Hu15) and 2.141.026 nt (Hu17) in 175 (Hu15) and 200 (Hu17) contigs. The main question here was how these two genomes differ from each other, especially concerning genes involved in penicillin resistance. This chapter shows details of this work which had not been included in the publication.

3.2.1 Genome comparison

3.2.1.1 Regions of divergent sequence

The comparison of the *S. pneumoniae* Hu15 and Hu17 genome sequences revealed several regions that differed between the two genomes. These regions consist of sequences present in only one genome or in both but with massive differences originating from exchanges or a high density of SNVs. Since the Illumina technology was used, the sequence quality at contig edges does not decrease and homopolymer stretches pose no problem and thus filtering of the alignment results is not as strict as described for 454 data in chapter 3.1. After removing regions from analysis which are located in gaps (one of the compared sub-sequences contains only N) as described at the analysis of the ST10523 genomes in chapter 3.1.1.1, ten regions were analysed in detail (see supplementary table S4). As also described in chapter 3.1.1.1, sequence gaps occur not always at the same position in the genomes and therefore some genes are missing in the overall analysis. The genome of Hu15 seemed to have a rearrangement of the sequences at the location 66.277 – 205.350 and 1.542.457 – 1.659.837 (corresponding regions in Hu17: 1.522.723 – 1.662.135 and 66.343 – 184.853) including the genes from SPNHU15_00077 to SPNHU15_00213 and from SPNHU15_01637 to SPNHU15_01767. Compared with the reference strain Hu^{19A}-6 (NC_010380) and with 454-sequence data (unpublished), this appears to be a mis-assembly. The apparently interchanged

sequence regions are directly flanked by gaps, which separate them clearly from preceding or subsequent sequence regions. Thus, there are no contigs containing sequence of these regions together with flanking regions of the first or second position. At the other hand, such contigs can be found in the unpublished 454-data, indicating a mis-assembly of the Illumina-data due to short and ambiguous reads.

Five of the analyzed regions contain a high SNP density, indicating recombination events. Indeed, three of them contain the genes encoding the penicillin-binding proteins Pbp2b, Pbp2x and Pbp1a and which are known to have a mosaic structure in penicillin-resistant strains. The comparison of these genes with *S. pneumoniae* R6 reveals almost identical sequences between R6 and Hu15, i.e. there was no indication of a mosaic structure, whereas Hu17 clearly contained mosaic PBP genes (Schweizer, et al., 2017). Two regions also contain a high density of SNPs and contain mainly genes of membrane associated proteins and several other proteins apparently not involved in penicillin-resistance. Another five of the analyzed regions affect only one single gene each. The insertions, deletions and replacement of these short regions are not likely to have arisen by repetition of sequence at contig edge or within a repeat. These inserted or deleted sequence fragments lead to frameshifts in the predicted genes and have to be verified by direct sequencing of this regions.

In summary, several differences in gene content between the two genomes are apparent, arisen by diverging sequence fragments (not SNVs). 71 (Hu15) and 75 (Hu17) genes from 2.151 (Hu15) and 2.157 (Hu17) genes, which were compared between the two genomes, were affected by insertions, deletions and exchanges as well as by high SNP density (likely due to recombination events). These genes represent three percent of the analysed genes.

Based on the publication of the genome sequences of *S. pneumoniae* Hu15 and Hu17, the penicillin-binding proteins PBP1a, 2b and 2x and the effect of their differences were further analysed and described by Schweizer *et al.* (Schweizer, et al., 2017). The proteins MurM and CiaH, mutations of which are also associated with penicillin-resistance and were described in this publication, were identical in Hu15 and Hu17. Thus, all three PBPs (1a, 2b, 2x) of Hu15 contain no mosaic blocks, in agreement with the penicillin-sensitive phenotype of Hu15. The presence of a *ciaH* mutation and a mosaic *murM* in Hu17, both associated with a penicillin-resistance phenotype, not only in Hu17 but also in the penicillin-sensitive strain Hu15, were surprising.

The plasmid (related to *pSpn1* as mentioned in the paper) which is present in each of the two strains is identical in both strains except for gaps by assembly.

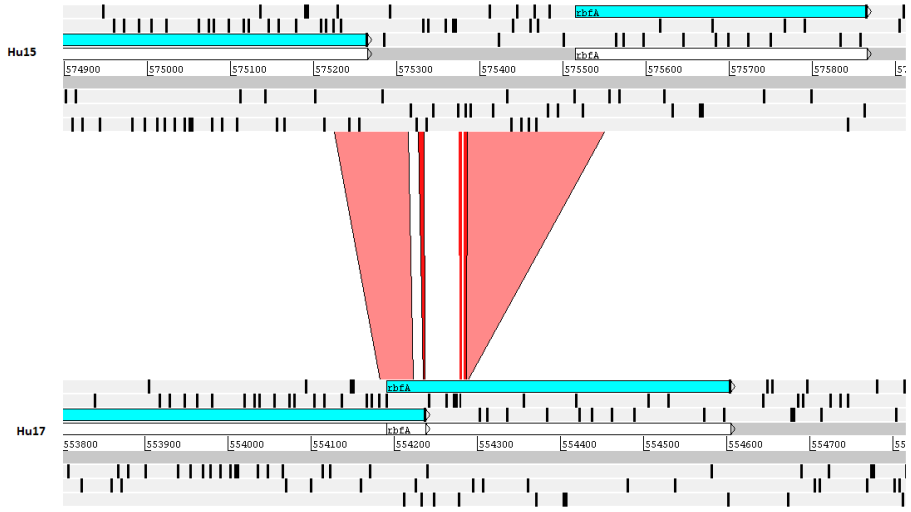


Figure 3.10: Apparent region exchange and deletions in the genes SPNHU15_00614 and SPNHU15_00615

Visualization with the Artemis comparison tool (ACT) shows the alignment of *Streptococcus pneumoniae* Hu15 (**top**) and Hu17 (**bottom**) at the location of the genes SPNHU15_00614 and SPNHU15_00615 and their counterparts SPNHU17_00609 and SPNHU17_00610. The first difference is an exchange of 89 nt in the genome of Hu15 by 40 nt in Hu17. 14 nt downstream of the exchanged region Hu15 contains four nucleotides (TTTG), which are absent in Hu17. 65 nt downstream of the exchanged region a sequence of 163 nt is missing in Hu17, which is present in Hu15. Two single nucleotide deletions (missing nucleotide in Hu17) lead up to this deletion.

3.2.1.2 Comparison of *S. pneumoniae* Hu15/Hu17with the closely related clone Hu^{19A}-6

In order to see whether the genomes of Hu15 and Hu17 contain specific genes, a comparative analysis was first performed with the closely related strain *S. pneumoniae* $Hungary^{19A}$-6 (Acc. No. NC_010380; afterwards referred to as Hu19A which represents a single locus variant (SLV) of the same clone), based on protein coding genes (CDS) filtered as described before. RNA genes were excluded for reasons described earlier in this work. 2.151 proteins of Hu15 and 2.157 proteins of Hu17 were compared with proteins of Hu19A and, as also described before, were considered as present with a coverage of at least 60% and an identity of at least 70%.

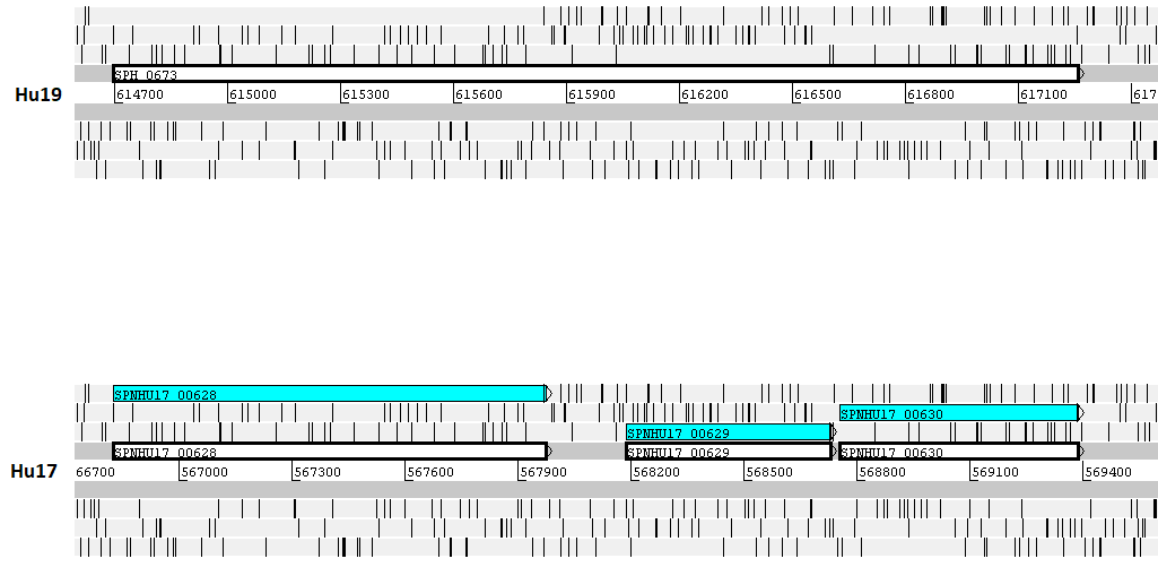


Figure 3.11: Representation of a gene with denoted authentic frameshift in Hu19

The visualization with the Artemis comparison tool (ACT) shows an example of the alignment between *S. pneumoniae* Hungary19A-6 (Hu19, **top**) and Hu17 (**bottom**) genome sequences. In Hu19 for 34 genes no CDS were annotated, since the genes might have internal stop codons, but the protein fragments were also not annotated. The proteins or protein fragments of Hu17 had to be compared separately with the deduced protein fragments of Hu19.

Hu19A contains 34 genes with annotated authentic frameshifts (Figure 3.11). Since there is no CDS annotated for these genes, they were manually compared with the proteins of Hu15 and Hu17.

The comparison resulted in the deduced proteins of 25 genes in Hu15 and 23 in Hu17 (supplementary table S5). 22 of these proteins were identical in Hu15 and Hu17. The genes SPNHU15_00707 (encoding a sodium/hydrogen exchanger family protein; Na+/H+ antiporter), SPNHU17_00868 (MutT/nudix family protein) and SPNHU15_01161 (hypothetical

protein) were present in Hu19A, which indicates frameshifts at positions, where the required coverage and identity could not be reached by the resulting protein fragments. Extension of the protein search at the NCBI homepage showed that all 23 Hu15 and Hu17 proteins are present in other *S. pneumoniae* strains. Only the proteins encoded by SPNHU15_00868/SPNHU17_00868 and SPNHU15_01681/SPNHU17_00123 could not be found. However, their DNA sequences were present in *S. pneumoniae* 670-6 and CSGP14 but were not annotated.

In summary, 98.84% respectively 98.93%, of the proteins of Hu15 and Hu17 are present in Hu19A in agreement with their clonal relatedness. The remaining proteins respectively their genes were present in other clonally unrelated *S. pneumoniae*. The penicillin-resistance determinants *murM* and *ciaH* are identical in all three genomes. In contrast, the PBP-encoding genes and flanking regions differ between the penicillin-resistant strains Hu17/Hu19A and the penicillin-sensitive Hu15 which resemble those of the laboratory strain R6, indicating functional differences.

3.2.2 SNVs in ST226

A detailed retrieval of single nucleotide polymorphisms (SNPs) including single nucleotide deletions and insertions (indels), was performed by manual comparison of the genes of the two ST226 genomes. This manual inspection confirmed the results of a test run of the software described in chapter 3.6. As mentioned above, not considered in this analysis was a subset of genes such as IS elements, repetitive elements and RNA coding and incomplete genes. Differences in homopolymer stretches were not omitted since this problem (described in chapter 1.2.3) does not occur with the Illumina sequencing technology. Furthermore, genes with differences located close (≤350 nt) to contig edges were also used in the analysis, in contrast to the comparison described in chapter 3.1.1.2, since the sequence quality at contig edges was much higher due to high coverage by reads. 64 - 66 CDS were excluded from analysis and 37 SNVs in 18 genes were further analysed (results are listed in detail in supplementary table S6). Six of these genes contain only silent SNPs. Most other genes containing amino acid exchanges and frameshifts seem not remarkable, except for two genes encoding ribosomal proteins: S12 (RpsL) and S6 (RpsF). RpsL is known to be involved in high-level streptomycin-resistance (Salles, et al., 1992). In the current analysis, a SNP results in a stop codon in *S. pneumoniae* Hu17 and thus to a length reduction of the amino acid sequence by ten amino acids (aa). Changes in length by extension at the N-terminus with effect on tRNA-binding behaviour could be observed in *Escherichia coli* (Calidas, et al., 2014). RpsF as a component of the 40S ribosomal subunit plays a crucial role in controlling cell survival and proliferation (Babina, et al., 2015). In *E. coli* and other bacteria, RpsF was observed to form heterodimers with the ribosomal protein S18 (RpsR) (Babina, et al., 2015). These heterodimers inhibit the translation of RpsF (Babina, et al., 2015). However, it was described as non-essential (Bubunenko, et al., 2007). The gene encoding RpsF (S6) contains a frameshift in *S. pneumoniae* Hu15/Hu19A and *S. pneumoniae* Hu17, leading to different 15 respectively 10 N-terminal amino acids. Such variants were not found in genome sequences of *S. pneumoniae* or other streptococcal species listed in the NCBI database by *BLAST* search. The SNVs should be verified by manual sequencing.

3.3 Analysis of streptococcal species isolated from different host organisms

The streptococcal genomes described in the publication originate from isolates obtained from primate and human hosts. Especially members of the species *S. oralis* which are part of the commensal human oral flora, can also be found in monkeys. Therefore, the analysis of genes and proteins present in *S. oralis* genomes obtained from different hosts was expected to reveal host specific components. Fortunately, the complete genome of one *S. oralis* strain Uo5 is available (Reichmann, et al., 2011). This chapter shows details of this work which had not been included in the publication. The generation of the scaffold sequences is based on 88.373 – 208.626 sequence reads with 15.968.893 – 40.461.884 nt obtained with 454 sequencing technology (supplementary table S1). 13.410.689 - 35.085.231 of reads were assembled using paired-end information into 25 - 2.883 contigs with a total of 1.672.711 - 3.031.270 nt (supplementary table S7).

3.3.1 Comparison of *S. oralis* genomes

Eleven of the genomes described in the publication could be assigned to *S. oralis*, nine (DD05, DD14-17, DD20-21, DD24-25) obtained from primates and two (DD27, DD30) from humans. In order to see which protein coding genes are common in members of this species obtained from different host organisms a comparison was performed, based on the reference strain *S. oralis* Uo5. After removal of genes encoding transposases, deduced proteins of 1.896 protein coding genes of *S. oralis* Uo5 were used for comparison with the eleven genomes. This comparison was performed as best hit retrieval with *TBLASTN* (Gertz, et al., 2006), to search the deduced *S. oralis* Uo5 proteins within

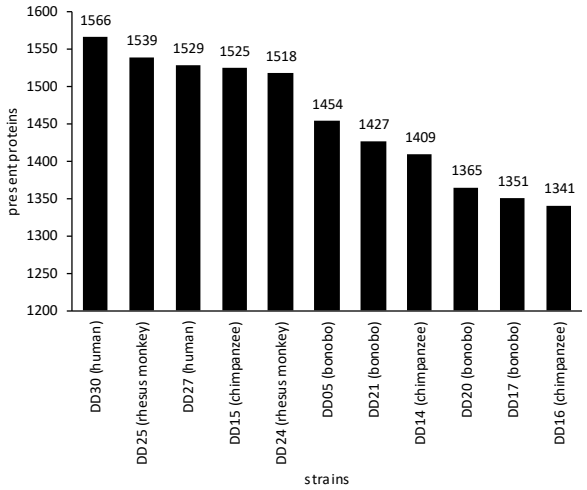


Figure 3.12: S. oralis Uo5-proteins present in S. oralis strains obtained from primates and human

Based on a coverage of ≥60% and an identity of ≥70%, eleven *S. oralis* strains obtained from primates and humans contain between 71 - 83% of the 1.896 deduced proteins of *S. oralis* Uo5, corresponding to 1.341 proteins of DD16 and up to 1.566 proteins of DD30. DD30 and DD29 as well as Uo5 originate from human hosts.

the nucleotide sequences of the eleven genomes contigs. It should be considered, that the contig sequences might contain errors in homopolymer stretches due to possible errors of the sequencing technology as described in the introduction. Also, matching sequences might be part of genes, which are incomplete at contig edges in the analysed genomes. Proteins were defined as present with at least 60% coverage and 70% identity as described above. 823 (44%) of the deduced *S. oralis* Uo5 proteins could be found in the eleven genomes, between 1.341 (71%) and 1.566 (83%) proteins in single genomes (Figure 3.12, supplementary table S8). The genomes of strains obtained from human hosts as well as the rhesus monkey strains and one chimpanzee strain contained 80 - 83% of the *S. oralis* Uo5 proteins. The second group of *S. oralis* genomes contained only 71 - 77% of *S. oralis* Uo5 proteins and were isolated from bonobo and chimpanzee. It should be kept in mind, that the rhesus monkeys had contact with humans as mentioned in the publication. In contrast, close contact of human to chimpanzees of the Thaï national park was not allowed, and strains were obtained from fruit residues (three genomes were obtained from *S. oralis*). Since random occurrence of the observed number of shared proteins in DD15 is not likely, evolution might have contributed and has to be investigated in detail. Incomplete genes (and accordingly proteins) as well as possible errors in homopolymer stretches were not removed from the analysis, values obtained should be taken as an approximation since not all genes were covered.

3.3.2 Genomes with pilus islet 2

As described in the publication, the pilus islet 2 (PI-2) (Zähner, et al., 2011) was identified in six genomes of streptococci from primates by screening for the deduced putative pilus backbone protein PitB of *S. oralis* Uo5. In addition, the PI-2 islet was also present in two *S. oralis* genomes: DD25 (rhesus monkey) and DD27 (human).

The PI-2 islet of DD25 is reduced in length due to the absence of the genes SOR_1070 (*pitA*), SOR_1069 (*sipA*), SOR_1068 (*pitB*) and the start of SOR_1067 (*srtG1*) (Figure 3.14). Since *pitA* and *pitB* encode the structural pilus proteins and *sipA* encodes a peptide essential for pilin synthesis (Zähner, et al., 2011), the PI-2 pilus is likely to be not expressed in *S. oralis* DD25.

Table 3.3: Presence of pilus proteins in DD27

The values represent coverage and identity of deduced proteins of the pilus islet 2 (PI-2) of DD27 compared to the five deduced proteins of the *S. oralis* Uo5 PI-2. The gene encoding pilus protein PitA is disrupted in DD27 but not by possible detection errors located in homopolymer stretches. Using 60% coverage and 70% identity as threshold, all PI-2 proteins are present in DD27.

locus_tag	gene	coverage	identity
SOR_1066	*srtG2*	100,00%	95,86%
SOR_1067	*srtG1*	100,00%	99,70%
SOR_1068	*pitB*	96,44%	78,08%
SOR_1069	*sipA*	100,00%	99,45%
SOR_1070	*pitA*	63,62%	84,70%
		37,89%	92,66%

The PI-2 cluster of *S. oralis* DD27 appeared intact, but with a large number of SNVs and indels and a disruption of the gene encoding PitA (Figure 3.13). This disruption appears not to originate from possible detection errors in homopolymer stretches. Comparison of the sequences of the *S. oralis* Uo5 and the *S. oralis* DD27 PI-2 proteins reveals the presence of all proteins with at least 60% coverage and 70% identity (Table 3.3). The first fragment of the disrupted protein PitA exceeds this threshold and thus, the protein is considered as present.

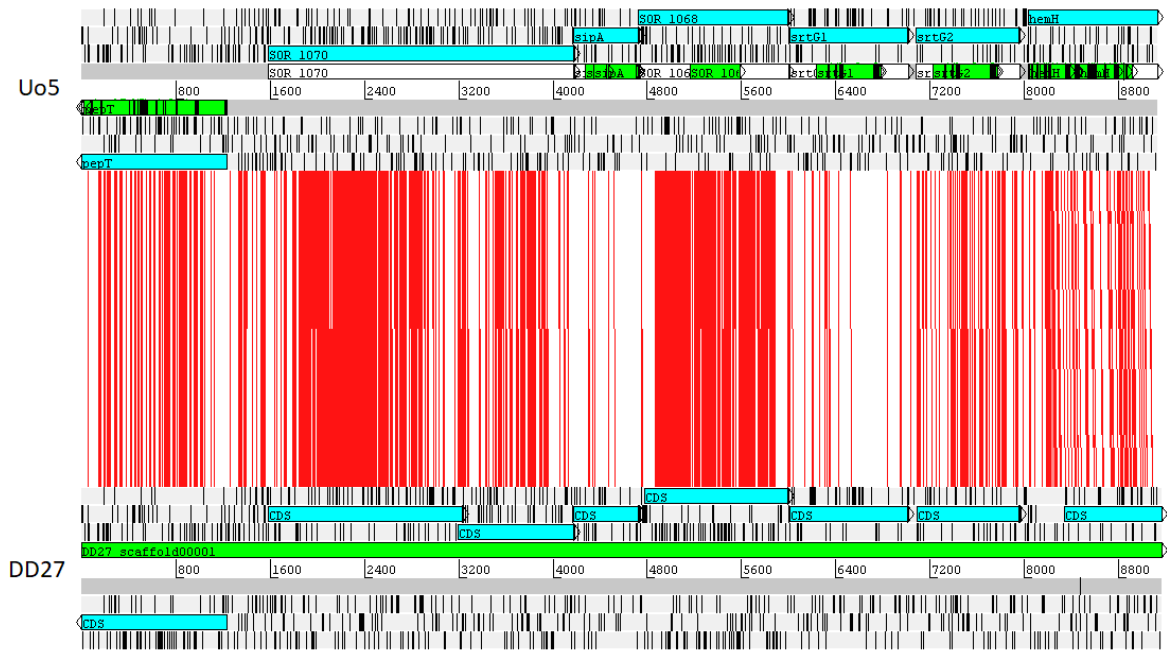


Figure 3.13: Comparison of the pilus islet 2 of S. oralis Uo5 and DD27

The visualization with Artemis comparison tool (ACT) shows the alignment of the pilus islet 2 (PI-2) of *S. oralis* Uo5 (**top**) and DD27 (**bottom**). Although massive differences at nucleotide level are apparent, all deduced proteins are present (Table 3.3). SOR_1070 is disrupted in DD27 but not by possible detection errors located in homopolymer stretches.

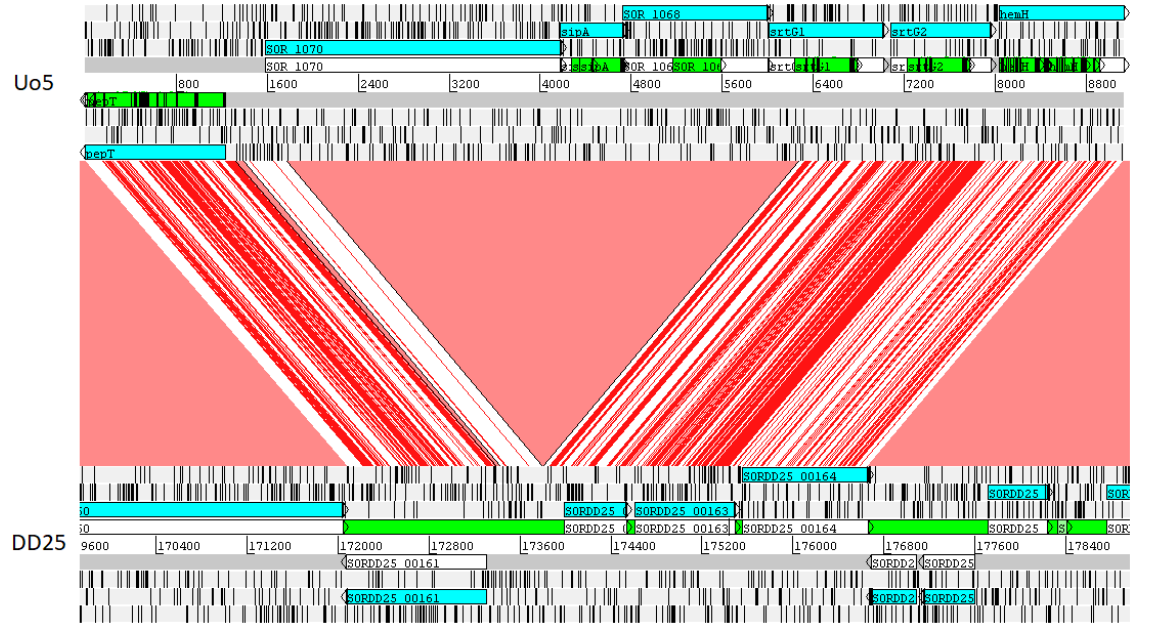


Figure 3.14: Genes absent in pilus islet 2 in DD25

The visualization with Artemis comparison tool (ACT) shows the alignment of the pilus islet 2 (PI-2) of *S. oralis* Uo5 (**top**) and DD25 (**bottom**) where three genes are absent in DD25. The three genes SOR_1068-70 encode the two pilus proteins PitA and PitB and the protein SipA, which is essential for pilin biosynthesis. Furthermore, the start region of the gene encoding the sortase SrtG1 is absent in DD25.

3.4 Analysis of *S. pneumoniae* R6 transformants obtained with *S. oralis* Uo5 DNA

Using *S. pneumoniae* R6 as recipient and DNA of *S. oralis* Uo5 and different beta-lactam antibiotics for selection, a series of high-level beta-lactam resistant transformants was generated. As described in the publication, the transformants were selected with piperacillin (P), cefotaxime (C) and then again with piperacillin and therefore named PCP. The genome of the transformant PCP-7 was sequenced and assembled, as well as the genomes of transformants obtained in two subsequent transformation and selection steps: PCP-C6 and PCP-CCO (selection with cefotaxime (C) and oxacillin (O)). To see whether and which differences occurred during these transformation steps, the genomes of these three transformants - PCP-7, PCP-C6 and PCP-CCO - were compared to each other, to the donor *S. oralis* Uo5 and the recipient *S. pneumoniae* R6 as described in this chapter. These transformants were already analysed in detail concerning β-lactam-resistance caused by recombination of penicillin-binding proteins and MurE (Todorova, 2010) using single gene sequencing, MIC (minimal inhibitory concentration) and microarrays. Genome-wide differences between these transformants based on the work described here were analysed in detail (Meiers, 2015). The data presented here were basic to the work of Todorova (Todorova, 2010; Todorova, et al., 2015) and Meiers (Meiers, 2015).

3.4.1 Generation of genome sequences

With Illumina (Hillier, et al., 2008; Liu, et al., 2012; Bentley, et al., 2008) sequencing technology sequence reads of the three transformants *S. pneumoniae* PCP-7, PCP-C6 and PCP-CCO were generated. This generation is based on 2.330.466 – 2.358.666 sequence reads with 351.900.366 – 356.158.566 nucleotides obtained (supplementary table S1). 309.345.207 (Hu15) and 307.838.725 nt (Hu17) of reads were assembled using paired-end information into 2.136.165 (Hu15) and 2.141.026 nt (Hu17) in 175 (Hu15) and 200 (Hu17) contigs. 330.220.372 (PCP-7), 328.527.136 (PCP-C6) and 335.158.827 nt (PCP-CCO) of reads were assembled into 1.987.196 (PCP-7), 1.987.398 (PCP-C6) and 1.987.327 nt (PCP-CCO) in 170 (PCP-7), 146 (PCP-C6) and 157 (PCP-CCO) contigs. The three sets of read data were assembled with *Newbler* (*gsAssembler*) (Margulies, et al., 2005) and then aligned with the genome sequence of *S. pneumoniae* R6. The generated genome sequences (unpublished) contain 1.971.219 nt

(PCP-7), 1.973.754 nt (PCP-C6) and 1.977.000 nt (PCP-CCO). The annotation of genes, CDS and RNA – genomic features - was manually transferred from *S. pneumoniae* R6 and *S. oralis* Uo5 to the PCP-transformants depending on alignments (*BLASTN*). In a first step, exchanged sequence regions were identified. The annotation of genomic features of *S. pneumoniae R6* was then transferred to the new genome at sequence regions, which are nearly identical (except for SNVs) to *S. pneumoniae* R6, otherwise the annotation of *S. oralis* Uo5 was used. During this process, transferred regions as well as SNVs were identified.

3.4.2 Genome comparison

3.4.2.1 *Transferred regions*

Complete genomes of the recipient *S. pneumoniae* R6 and of the donor *S. oralis* Uo5 were available (references), and therefore recombined regions in the transformants could be identified unambiguously. The alignment of the generated sequence contigs of the transformant PCP-7 with the *S. pneumoniae* R6 genome revealed nine regions where the alignment failed due to the presence *of S. oralis* Uo5 sequences. All nine *S. oralis* Uo5 regions were also present in PCP-C6 and PCP-CCO (Table 3.4) as described by Meiers (Meiers, 2015). Four regions contained sequences of *S. oralis* genes encoding Uo5 Pbp2x, Pbp2b, Pbp1a and surprisingly MurE and contributed to the increased resistance of the transformants. The region with the gene encoding Pbp1a contained a short (238 nt) fragment of *S. pneumoniae* R6 sequence within the RecU gene with a silent SNP, indicating two closely neighboured or nested recombination events. The same region also contains an intergenic indel. In addition, five regions with apparent recombination events contained genes encoding hypothetical proteins, ABC-transporter components, a topoisomerase, a tRNA-synthetase, a deacylase and an integrase. Apart from the mentioned differences, the transferred regions were identical to the donor sequence. Surprisingly, the genomes of PCP-C6 and PCP-CCO showed no further recombination events, but one SNP within the exchanged region in the *pbp2b* gene resulted in the change of Gln406 into Pro in the *S. oralis* Uo5 sequence. These findings were confirmed by additional manual sequencing (Meiers, 2015).

Table 3.4: Transferred regions in PCP-genomes

The genome of *S. pneumoniae* PCP-7 contains nine recognizable regions of *S. oralis* Uo5 sequence, where one region contains a short fragment of *S. pneumoniae* R6 sequence (within gene encoding RecU) besides an intergenic SNP. In the subsequent transformants PCP-C6 and PCP-CCO do not show further sites of apparent recombination. Only the penicillin binding protein 2b of PCP-CCO is affected by a SNP, which changes an encoded glutamine at position 406 into a proline.

R6			PCP-7			PCP-C6			PCP-CCO				
start	stop	length	start	stop	length	start	stop	length	start	stop	length	affected genes (homologue of *S. pneumoniae* R6/*S. oralis* Uo5)	product
155.942	158.292	2.351	154.921	157.290	2.370	155.942	158.311	2.370	155.942	158.311	2.370	spr0147/sor1832[1]	ABC transporter solute-binding protein - unknown substrate
												sor1831	putative deacylase
												spr0149/sor1830[1]	ABC transporter ATP-binding protein - unknown substrate
303.118	305.240	2.123	303.436	305.558	2.123	303.583	305.705	2.123	303.137	305.259	2.123	spr0304/sor0341[1]	Penicillin-binding protein 2X
												spr0305/sor0342[1]	Undecaprenyl-phosphate-UDP-MurNAc-pentapeptide phospho-MurNAc-pentapeptide transferase MraY
332.522	335.505	2.984	332.541	335.547	3007[4]	332.541	335.547	3007[4]	332.541	335.547	3007[4]	spr0328/sor1642[1]	hypothetical protein
												sor1641	penicillin-binding protein 1A
												spr0330/sor1640[1]	Recombination protein U (RecU)[3]
749.125	751.747	2.623	748.689	751.311	2.623	748.689	751.311	2.623	748.689	751.311	2.623	sor0818	hypothetical protein
												spr0756/sor0819[1]	Topoisomerase IV subunit B
931.672	931.775	104	931.236	931.339	104	931.236	931.339	104	931.236	931.339	104	spr0947/sor1025[1]	hypothetical protein
1.040.033	1.040.868	836	1.039.597	1.040.432	836	1.039.597	1.040.432	836	1.039.597	1.040.432	836	spr1046/sor0993[1]	Integrase/recombinase
1.366.477	1.369.803	3.327	1.366.201	1.369.522	3.322	1.366.246	1.369.568	3.323	1.366.075	1.369.396	3.322	spr1383/sor0660[1]	polysaccharide transporter
												sor0659	UDP-N-acetylmuramyl tripeptide synthase MurE
												sor0658	hypothetical protein
												sor0657	hypothetical protein
1.480.794	1.482.807	2.014	1.480.387	1.482.400	2.014	1.480.517	1.482.530	2.014	1.480.387	1.482.400	2.014	spr1502/sor0586[1]	isoleucyl-tRNA synthetase
1.494.914	1.497.390	2.477	1.494.507	1.497.027	2.521	1.494.637	1.497.157	2.521	1.494.507	1.497.027	2.521	spr1517/sor0561[1]	Penicillin-binding protein 2B[2]
												sor0560	phosphosugar-binding transcriptional regulator
												spr1519/sor0559[1]	Glucokinase

[1] gene contains sequence fragments of donor and recipient
[2] SNP in PCP-CCO: $A^{7/C6}{}_{1217}C^{CCO}$ ($Q^{7/C6}{}_{406}P^{CCO}$)
[3] exchanged region contains R6-sequence at location 49-286 (position 159-396 in gene, contains one SNP: $A^{R6}{}_{192}C$ (silent))
[4] region contains an intergenic indel: $TT^{Uo5}{}_{1596740/c332861}T$

3.4.2.2 SNPs and other differences

Besides the apparent regions of recombination described before, some SNPs, single and short sequence indels appear in the three transformants, compared to R6 and each other (Table 3.5). Differences in incomplete genes, gaps or repeats were ignored.

In the genome of PCP-7, three SNPs occurred compared to R6, which are also present in the two subsequent transformants: An intergenic SNP and amino acid changing SNPs in spr0087 (encoding a hypothetical protein) and spr0738 (encoding the purine nucleoside phosphorylase DeoD).

In the genome of PCP-C6, two SNPs appeared compared to PCP-7, which were also present in the genome of PCP-CCO: one non-silent SNP in the gene spr1992 (encoding a hypothetical protein) and another one within the gene spr0708 encoding the histidine protein kinase CiaH, where mutations frequently affect penicillin susceptibility (Müller, et al., 2011; Meiers, 2015).

PCP-7 and PCP-C6 contain further differences to *S. pneumoniae* R6 respectively PCP-7, which occurred only in the analysed genome and not in subsequent transformants and therefor were ignored in the analysis of penicillin resistance determinants.

Based on the observation described above and because there is no genome sequence of a potential subsequent transformant available, it is not easy to decide, which alteration in the genome of PCP-CCO is authentic. This genome contains, compared to the preceding transformants, one intergenic SNP and two intergenic indels, moreover a deletion of 157 nt within the gene spr0415 (encoding the pyruvate formate-lyase Pfl), a deletion of 66 nt at the end of spr1835 and the start of spr1336 (encode the cellobiose-specific IIB component PtcB and IIA component PtcA of a phosphotransferase system) and a deletion of ten nucleotides within the gene spr2045 (encodes the serine protease Sphtra), which is involved in competence control (Schnorpfeil, et al., 2013; Laux A, 2015) and is known as major virulence factor (Ibrahim, et al., 2004). The deletion results in a truncated 124 aa protein product.

In summary, there are only few differences between the three transformants and to *S. pneumoniae* R6. As described above, there is no apparent recombination after PCP-7. The transformants contain only one or two amino acid changing SNPs within genes. Furthermore, there are two to five genes containing deletions of more than one nucleotide. As stated in this chapter, these deletions have to be verified, except for the deletion in the Sphtra gene, which

was confirmed by an alternative sequencing method (Meiers, 2015). The differences listed here are described in detail by Meiers (Meiers, 2015).

Together, these studies revealed several important issues. First, *mur*E was identified as a penicillin-resistance determinant. Second, it became clear that it is not possible to transfer the entire resistance potential phenotypically expressed in *S. oralis* Uo5 into *S. pneumoniae*. Third, the contribution of CiaH, HtrA and Pbp2b alleles to beta-lactam resistance was evident.

Table 3.5: Differences between PCP sequences and S. pneumoniae R6

During the underlying transformation series several SNPs, indels and region indels occurred. Apart from recombined regions, the genome of PCP-7 contains three SNPs compared to R6, which also are carried by the other two transformants. Two further SNPs appear in the genome of PCP-C6: spr0708 (*ciaH*), $A_{258}E$; spr1992 (hypothetical protein), $I_{272}T$. In the genome of PCP-CCO three intergenic SNPs and indels came along with three deletions of short sequences within genes. PCP-7 and PCP-C6 in each case additionally contain seven SNPs, indels or short sequence indels compared to R6, which are not present in the subsequent transformants.

R6			PCP-7		PCP-C6		PCP-CCO	
location	locus_tag	product	location	difference	location	difference	location	difference
94.957	spr0087	hypothetical protein	92.954	$G_{852}A$ ($R_{284}K$)	94.426	$G_{852}A$ ($R_{284}K$)	94.957	$G_{852}A$ ($R_{284}K$)
139.371 - 139.375	spr0131	Secreted metalloendopeptidase Gcp	-	no difference to R6	139.463	$TTTGG_{381}G$	-	no difference to R6
139.388 - 139.476	spr0131	Secreted metalloendopeptidase Gcp	-	no difference to R6	139.476	deletion of 88 nt at position $398^{R6}/394^{PCP\text{-}C6}$	-	no difference to R6
412.615 - 412.274	spr0415	Pyruvate formate-lyase Pfl	-	no difference to R6	-	no difference to R6	412.659	deletion of 157 nt at position 503
450.589	-	intergenic	449.730	$C_{450589/449730}T$	449.804	$C_{450589/449804}T$	450.474	$C_{450589/450474}T$
548.623	spr0547	dipeptidase PepV	-	no difference to R6	547.952	$GC_{595}C$	-	no difference to R6
548.630 - 548.688	spr0547	dipeptidase PepV	-	no difference to R6	547.958	deletion of 58 nt at position $601^{R6}/600^{PCP\text{-}C6}$	-	no difference to R6
708.729	spr0708	sensor protein CiaH histide kinase	-	no difference to R6	708.528	$C_{773}A$ ($A_{258}E$)	708.293	$C_{773}A$ ($A_{258}E$)
736.926	spr0738	purine nucleoside phosphorylase (inosine phosphorylase) DeoD	736.490	$G_{421}T$ ($A_{141}S$)	736.692	$G_{421}T$ ($A_{141}S$)	736.490	$G_{421}T$ ($A_{141}S$)
820.955	-	intergenic	-	no difference to R6	-	no difference to R6	820.519	$G_{820955/820519}C$
1.210.035	-	intergenic	-	no difference to R6	-	no difference to R6	1.209.491	$GA_{1210035/1209491}A$
1.210.060	-	intergenic	-	no difference to R6	-	no difference to R6	1.209.515	$C_{1210060/1209515}TC$
1.226.163 - 1.226.285	spr1228	Shikimate kinase AroK	1.225.741	deletion of 58 nt at position 59	-	no difference to R6	-	no difference to R6
1.266.586 - 1.266.627	spr1272	N-acetylglucosamine-6-phosphate isomerase NagB	-	no difference to R6	1.266.280	deletion of 41 nt at position 433	-	no difference to R6
1.394.593 - 1.394.594	spr1410	putative calcium transporter PacL	1.394.195	$AC_{1851}C$[1]	-	no difference to R6	-	no difference to R6
1.394.808 - 1.394.816	spr1410	putative calcium transporter PacL	1.394.409	$AATCCGTCC_{1629}C$[1]	-	no difference to R6	-	no difference to R6
1.394.847 - 1.394.851	spr1410	putative calcium transporter PacL	1.394.440	$GCCGC_{1594}C$[1]	-	no difference to R6	-	no difference to R6
1.394.889 - 1.394.890	spr1410	putative calcium transporter PacL	1.394.478	$AC_{1555}C$[1]	-	no difference to R6	-	no difference to R6
1.394.912 - 1.395.036	spr1410	putative calcium transporter PacL	1.394.500	deletion of 124 nt at position 1409	-	no difference to R6	-	no difference to R6
1.428.321	-	intergenic	1.427.785	$C_{1427785}A$	-	no difference to R6	-	no difference to R6
1.984.498	spr1992	hypothetical protein	-	no difference to R6	1.985.059	$T_{815}C$ ($I_{272}T$)	1.984.069	$T_{815}C$ ($I_{272}T$)
1.807.546 - 1.807.649	spr1833	Beta-glucosidase Bgl2	-	no difference to R6	1.808.109	deletion of 103 nt at position 138	-	no difference to R6
1.810.420 - 1.810.486	spr1835/spr1836[2]	Phosphotransferase system system, cellobiose-specific IIB component PtcB and cellobiose-specific IIA component PtcA	-	no difference to R6	-	no difference to R6	1.810.057	deletion of 66 nt at position $243^{spr1835}/\text{-}45^{spr1836}$
2.023.817 - 2.023.916	spr2033	Inosine-5'-monophosphate dehydrogenase ImdH	-	no difference to R6	2.024.561	deletion of 99 nt at position 1.113	-	no difference to R6
2.036.726 - 2.036.738	spr2045	Serine protease Sphtra	-	no difference to R6	-	no difference to R6	2.036.085	deletion of 10 nt at position 333

[1] position in R6 allele due to several frameshifts in PCP-7
[2] missing sequence fragment contains end of spr1835 and start of spr 1836

3.5 Common genes of different streptococcal strains and species

In the book chapter, core genome analyses of representatives of the streptococcal species *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis* and *S. oralis* are described as well as the comparison of the *S. pneumoniae*-specific proteins with the proteins of 26 complete *S. pneumoniae* strains. This work visualizes differences as well as common features between species and strains, and the impact of horizontal gene transfer within and between species.

3.5.1 Comparison of individual streptococcal genomes representing different species

To examine the proteins common to *S. pneumoniae* and its close relatives, the deduced proteins of *S. pneumoniae* R6 (1.935 proteins) were compared to those of *S. mitis* B6 (1.937) and *S. oralis* Uo5 (1.898), excluding transposases and IS-elements. In addition, *S. mitis* B6 and *S. oralis* Uo5 specific genes were retrieved. In this context it is important to note that *S. pneumoniae* R6 is a penicillin sensitive laboratory strain isolated over 80 years ago, whereas both, *S. mitis* B6 and *S. oralis* Uo5, are multiple antibiotic and high-level penicillin resistant strains indicating several gene transfer events in the latter two strains. Using a threshold of 60% coverage and 70% identity for the definition of common deduced proteins (Denapaite, et al., 2010), a minimum of 1.140 proteins is present in all three species. This number differs slightly when pairwise comparisons are performed and does not necessarily represent common gene content, since some genes might be fragmented in one genome (i.e. the product is not present), whereas it is intact in the other. The pan-genome of these three species was determined to include 3.057 proteins/protein-coding genes. The estimated core of 1.140 proteins amounts to 37% of the pan-genome and 59 – 62 % of the individual genomes. The percentage of proteins shared by two genomes reflects the evolutionary relationship between the three species. *S. oralis* and *S. mitis* could be isolated in Old World monkeys held in captivity and are supposed to have evolved from a common ancestor prior to specialization of *S. pneumoniae* out of *S. mitis* in a common ancestor of primates and human (see reference book chapter, and chapters 2.3 and 3.3). According to MLST data as well as deduced from genomic comparisons, *S. pneumoniae* represents a specialized *S. mitis* lineage evolved in the human host (see book chapter). In agreement with this, *S. pneumoniae*

R6 shares more proteins with *S. mitis* B6 (1.321) than with *S. oralis* Uo5 (1.237). A total of 345 proteins (11 % of the pan-genome) is shared by only two of the three species (9 – 14 % per species). 1.572 proteins (51 % of the pan-genome, 26 – 30 % of individual genomes) are found to be specific to one of the three species.

Recently, the species *S. pseudopneumoniae* was defined, which is placed in a distinct group between *S. pneumoniae* and *S. mitis* according to MLSA (multi locus sequence analysis) data (see book chapter). Since the complete genome of *S. pseudopneumoniae* strain IS7493, accession number NC_015875) (Shahinas, et al., 2011) was available, we included this species in a final comparison. 1.105 deduced proteins were common to the four species, and 1.446 *S. pneumoniae* R6 proteins are present in *S. pseudopneumoniae* IS7493.

3.5.2 Global comparison of *Streptococcus pneumoniae* with other streptococcal species

To see, which of the 532 proteins defined as *S. pneumoniae* R6-specific in the comparison with *S. mitis* B6 and *S. oralis* Uo5 are common to *S. pneumoniae* in general and which are specific for this strain, 26 public available complete genomes of *S. pneumoniae* were used (supplementary table S9) using the same cut-off values as before. *S. pseudopneumoniae* was not included since it was not available during the first analysis; see previous paragraph. 104 proteins were present in all these genomes (supplementary table S10). 67 of their genes were organized in 17 clusters composed of two to twelve genes. They mainly represent transporters, proteins related to sugar-metabolism and hypothetical proteins as well as variants of the competence stimulating peptide precursor ComC (competence stimulating peptide (CSP), essential for competence (Laux A, 2015)) and the two-component system of HK06 and RR06, which is known to regulate expression of the major virulence factor choline binding protein A (CbpA), PspA and other proteins related with adhesion (Standish, et al., 2005; Standish, et al., 2007). ComC is also present in other genomes but with highly altered sequences, similar to HK06 and RR06.

This analysis was repeated after publication of the first *S. pseudopneumoniae* genome in order to further specific *S. pneumoniae* specific virulence factors (Figure 3.15).

All (filtered) 1.935 deduced *S. pneumoniae* R6-protein sequences were searched for presence in the genomes of *S. mitis* B6, *S. oralis* and *S. pseudopneumoniae* IS7493 as described in the previous chapter. The addition of *S. pseudopneumoniae* resulted in a reduced number of common proteins from 59% (1.140) to 57% (1.105) (Figure 3.15, second row) and *S. pneumoniae* R6-specific proteins from 532 (27%) to 384 (20%), while the number of proteins shared by any but not all species increased from 263 (14%) to 446 (23%). This result shows that the addition of one closely related species to the analysis affected the number of common proteins only slightly, whereas a significant decrease of species-specific proteins was observed.

Proteins specific for *S. pneumoniae* R6 proteins were further defined in a comparison with another finished 26 *S. pneumoniae* genomes (supplementary table S9). Only 1% (22) of the proteins of *S. pneumoniae* R6 was not found in any of the other genomes. These 22 proteins are hypothetical proteins, which can be found in other (incomplete) genomes at the NCBI

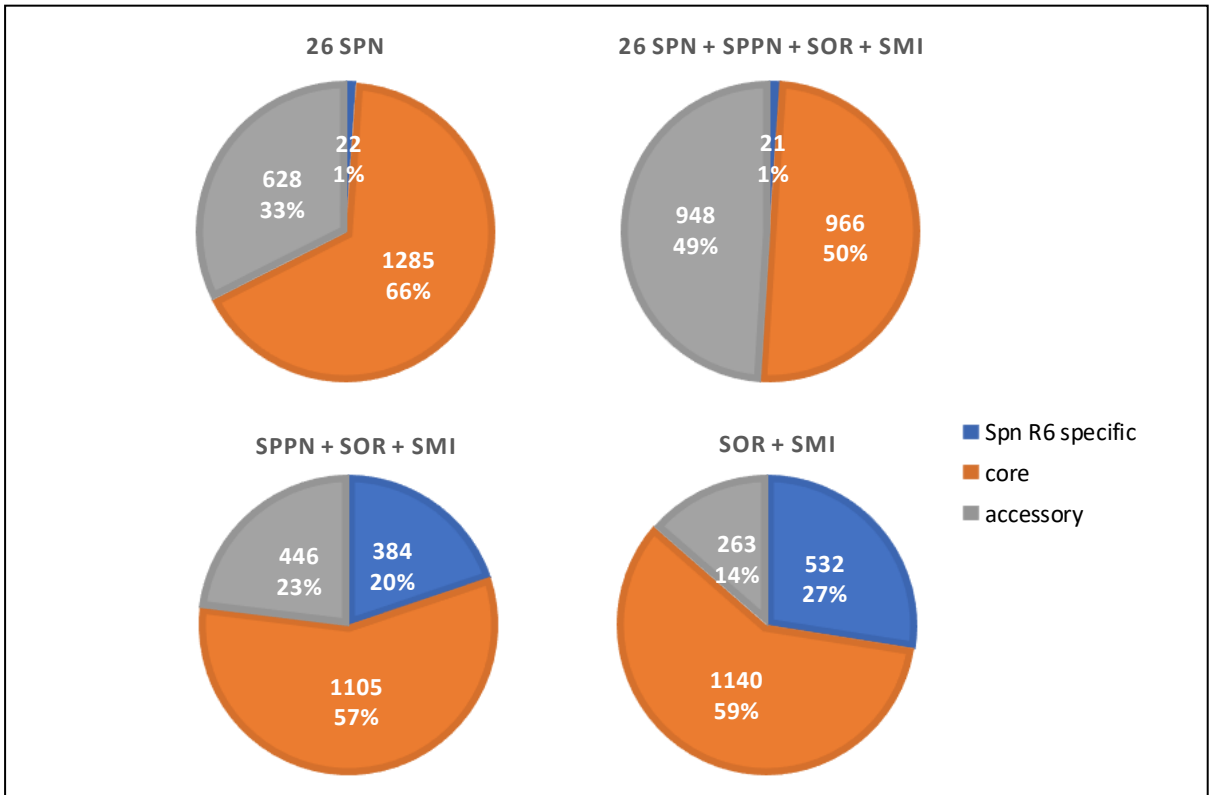


Figure 3.15: Overview of common and special proteins

Proteins of *S. pneumoniae* R6 were compared to deduced proteins of 29 publicly available *S. pneumoniae* genomes (29 SPN) and to those of representatives of closely related species (*S. mitis* B6 (SMI), *S. oralis* Uo5 (SOR) and *S. pseudopneumoniae* (SPPN)) using a minimum identity of 70% and a minimum coverage of 60%. The addition of the SPPN genome (right) affected mainly the species-specific features and the accessory genome whereas little effect on the number of common proteins was noted.

homepage. 66% (1.285) of the proteins are common to all genomes. This result deviates slightly from the results reported by Donati *et al*. (Donati, et al., 2010), who calculated that 74% of the DNA sequences were common to all genomes using the genomes of 14 complete and 30 incomplete *S. pneumoniae* genomes. The difference is probably due to the inclusion of incomplete genomes, as well as using DNA sequences as the basis for the comparison. Another study was based on the definition of orthologous clusters (Hiller, et al., 2007). Here, 17 pneumococcal genomes were used (seven of these genomes in addition to *S. pneumoniae* R6 are used in the current work), and 21 - 32% of all CDS (or orthologous cluster) of any *S. pneumoniae* genome were not associated with the core genome. 46% of the 3.170 analysed clusters were conserved among all genomes. The current work did not use orthologous clusters, but distinct deduced proteins of one reference genome (*S. pneumoniae* R6) and a similar number of non-core genes (34%) was identified (or proteins). Croucher *et al.* (Croucher, et al., 2013) analysed 616 mainly unfinished *S. pneumoniae* genomes. Their calculation was based on the definition of orthologous clusters, i.e. functional similar genes but not similar sequences, resulting in 1.194 orthologous clusters (out of a total of 5.442) present in a single copy in all genomes and thus representing the pneumococcal core.

In summary, all these studies confirm a large accessory genome of *S. pneumoniae* and other related streptococci as well.

The current analyses revealed 1.285 proteins common to 26 *S. pneumoniae* genomes, 966 of which were shared with all other related streptococcal genomes analysed.

And what differentiates *S. pneumoniae* R6 from other pneumococci? Since only a few proteins are left after analysis which are specific for this genome which can be found in other (incomplete) *S. pneumoniae* genomes, the individuality and abilities of *S. pneumoniae* R6 arise not from special genes or proteins but rather from individual point mutations and the absence of some (functional) genes.

3.6 Software development

The main questions concerning the current work are based on comparisons of genome sequences to retrieve differences of strains that belong to the same *S. pneumoniae* clone (strains from different patients that belong to the novel serotype 23F ST10523 clone, and strains varying in their antibiotic resistance pattern of the clone ST226, a single locus variant (SLV) of the multiple antibiotic resistant clone Hungary^{19A}-6), transformants of the laboratory strain R6 obtained with DNA of high-level resistant closely related oral streptococci, and genomes of different streptococcal species to identify species specific genes. Besides presence or absence of genomic islands, gene clusters or single genes, also single nucleotide variations (SNV) are important for the analysis of clones and transformants. There are several genome comparisons described for *S. pneumoniae* and other streptococci. For example, Croucher *et al.* (Croucher, et al., 2011) examined 240 *S. pneumoniae* genomes including serotype switch variants of the international important multiple resistant *S. pneumoniae* clone Spain^{23F}-1. Fernandes *et al*. (Fernandes, et al., 2017) compared five *S. pyogenes* isolates to the *S. pyogenes* MGAS5005 genome to reveal differences which can explain factors of invasive infections, and Wyres *et al.* (Wyres, et al., 2012) compared 426 pneumococcal genomes of several serotypes and of 70 years for a better understanding of evolution and penicillin-resistance in this species. However, differences in highly variable genes which are the result of gene transfer events were often not distinguished from true SNVs that originate by mutations, resulting in a distortion of e.g. phylogenetic analysis based on SNVs. There are several publicly available tools offering a broad variety for sequence alignment, analysis and visualization as well as for the conversion of file formats required for the input. For example, *BLAST* (Altschul, et al., 1990), *CLUSTAL* (Sievers, et al., 2014; Thompson, et al., 1994) and *Mauve* (Rissman, et al., 2009) are well known tools used for sequence alignments as *MEGA4* (Tamura, et al., 2007), *PHYLIP* (Felsenstein, 2013), *PAUP* (Swofford, 1999) and other are used for phylogenetic analyses. SNP analysis can be performed with tools like *GATK* (McKenna, et al., 2010), *SNPsFinder* (Song, et al., 2005) and *Mummer* (Delcher, et al., 1999; Delcher, et al., 2002; Kurtz, et al., 2004; Marçais, et al., 2018), visualization with *Artemis* (Carver, et al., 2012) or *ACT* (Carver, et al., 2005). But naturally, the borders of the mentioned categories are not fix for each tool. For example, SNP retrieval and other functions can also be performed by using BLAST or Artemis.

During the retrieval of SNVs and other differences such as insertions or deletions of larger regions of the ST10523 genome sequences, problems occurred due to the incompleteness – gaps within the sequences represented by stretches of N – of the compared sequences, and because there was no reference sequence for this clone. Empirically, available programs are not able to distinguish between variable and not variable genes. Furthermore, they do not meet further requirements for this analysis simultaneously as outlined below. Instead of sequence reads, the analysis software should allow input of genome sequences containing feature annotation in EMBL (European Molecular Biology Laboratory) file format if available and should be widely independent from preceding sequencing technology. Another important aspect is the form of output. It has to be human readable and, if possible, usable for visualization, particularly by the *Artemis* (Carver, et al., 2012) or *Artemis* comparison tool (*ACT*) (Carver, et al., 2005). Since the underlying data were generated with 454 sequencing technology, under- and overcalls in homopolymer stretches might occur and differences of the compared sequences in such stretches should be recognized. Also, the software should be able to mark or discard differences within a certain distance to contig gaps due to possibly decreasing sequence quality at contig edges. Furthermore, gaps of incomplete sequences have to be considered, and information of already annotated genes should be taken into account at least at output generation.

For example, the tool *SAMtools* (Li, et al., 2009) offer functions for data of several sequencing technologies from read manipulation to alignments to text-based visualization, but input sequences are sequence reads in MAQ (Li, et al., 2008) file format, which e.g. in case of 454 (SFF files) and Illumina (FASTQ format (Cock, et al., 2010)) have to be converted first. Many other programs could be listed offering parts of the desired functionality, especially approved alignment tools like *BLAST* or *Clustal* ((Sievers, et al., 2014; Thompson, et al., 1994)). The core principle of a further tool called *Wasabi* (Web Accessible Sequence Analysis for Biological Inference) (Kauff, et al., 2007) where personal experience is available (transfer of the source code into another programming language), which performs refinements of already existing multiple alignments, but not of two input sequences, was considered suitable for a basic workflow to compare two sequences in detail, wrapped by format converters to meet the requirement of a common input file format and a summarizing human readable and visualizable output. The program *Wasabi* itself was not used.

Wasabi uses multiple alignments in Nexus (Maddison, et al., 1997) format as input for refinement with block alignments. The output is also in Nexus format. Since this is not suitable for aligning two sequences in EMBL format and generating output files for visualization with *Artemis* or *ACT*, a new program was developed based on the core idea of *Wasabis* block alignment, but with changes concerning input and output format as well as alignment algorithm. Furthermore, SNV retrieval for aligned regions was integrated into the workflow as described below.

The developed program offers no new functionality but employs the established tool *BLASTN* (Altschul, et al., 1990) wrapped by pre- and postprocessing of the data. This enables the tool to add annotation information to the output, recognize SNVs as well as not alignable regions and generate output in tabular human readable form and files used for visualization with *Artemis* and *ACT*. Regarding the usage with 454 data, SNVs in homopolymer stretches are specifically marked for manual inspection but this might be ignored using data of other sequencing technologies. Since the annotation of genes showing differences is inherited into the results, it is possible to distinguish them e.g. for mobile elements, variable genes and so on. The program was developed and tested with pairwise comparisons of the genome sequences (containing stretches of N) of the three ST10523 isolates. Advantage of these genomes was their high sequence similarity, which provide a manageable and verifiable number of differences. This program was tested using the single gene cluster of the capsule locus of these genomes and the reference strain *S. pneumoniae* ATCC700669 (23F). Usage of the program with the ST226 genome and plasmid sequences, which were generated from shorter Illumina sequence reads, was also successful and emphasized the problem of diverging genomic arrangement. Success means, that detected differences were confirmed by detailed manual inspection. Comparisons of 454 sequences as well as of Illumina sequences regardless of the type of sequence (genome, plasmid, gene cluster) worked equally well and could be used for analyses of further sequences. In the context of an unpublished study concerning *S. pneumoniae* strains associated with meningitis, eight members (strains U22, 456, 496, 638, PS4401, PS184, F10, SA17) of the clone Spain^{23F}-1 were sequenced with 454 sequencing technology, and the comparison with the reference strain of the clone, *S. pneumoniae* ATCC700669 (also called *S. pneumoniae* 23F), represents another opportunity to use the

developed software for the investigation of SNVs to evaluate the evolution of clonally related strains.

3.6.1 Analysis software workflow

The workflow of the newly developed software, which is written in *Java* (Java) and requires *BLASTN* (Altschul, et al., 1990), *ACT* (Carver, et al., 2005) and *BioJava* (Prlic, et al., 2012) *1.5*, is visualized in Figure 3.16 and described as follows.

The program uses EMBL formatted files as input, where the presence of a sequence, even a gapped sequence, is important rather than gene annotation. Files containing more than one sequence entry are not allowed. DNA sequences were extracted from these files for further analysis and, if present, gene annotation for supplementary information in the output files.

The two DNA sequences then are searched with *BLASTN* (default parameters) for the best matching region. If such a region is found, the sequences are split into three fragments: The matching sequence pair and if available the sequence pairs located left and right of it. The search for the best matching region is repeated recursively for the left and right sequence pairs. If one of the pairs consist of only one sequence, because the other pair member is not existent, this is noted in the output as unaligned region. The same applies to pairs, where both sequences are present, but no matching region can be found by *BLASTN*. The matching pairs are aligned by a Needleman-Wunsch algorithm (Needleman, et al., 1970) provided by *BioJava* to determine SNPs and Indels (single nucleotide insertions or deletions), together referred to as SNV (single nucleotide variation), of the region. The determined SNVs are reported in the output files. In addition to the position and the kind of difference, the lower deviation of a configured threshold of distance to the next gap as well as the ten left and right flanking nucleotides, an indicator of possible homopolymer error and possibly present gene information is written into output. The homopolymer indicator is set at differences in homopolymer stretches of at least three identical nucleotides, where an insertion or deletion of at least one of the same nucleotides occurs.

The output of the analysis is divided into several types. The first set of output consists of tabular text files containing human readable information about unaligned regions, SNPs and Indels. For SNVs, if sequence annotation is available, gene location, locus tag, product and gene and protein sequences and lengths are added, if present.

Furthermore, files needed for visualization with *ACT* are generated. These contain tabular comparison files and executable batch files for SNVs as well as for unaligned, aligned and all

regions and SNVs together with unaligned regions. Execution of a batch file starts the visualization. This visualization enables the user to find differences more easily and a simple click into the sequence, genomic feature or difference of interest facilitates detailed inspection including usage of *ACT*-intern tools.

For these files filtered variants are generated except for aligned regions. The filter removes regions and SNVs, where one sequence contains only N (see chapter 3.1.1.1), since they represent gaps or ambiguous nucleotides. SNVs and regions, which are completely located within a given threshold of distance to a gap, are absent in the filtered data sets. Differences in homopolymer stretches remain, since the analysed sequences were not necessarily generated by 454 sequencing technology.

The program needs only a few minutes to perform the complete analysis procedure, but this duration is dependent on the similarity of compared sequences due to the number of detailed analysis steps and the hardware used: about 360 – 450 seconds for each comparison of the three ST10523 genomes on a machine with four GB RAM and 2x 3.07 GHz and about 140 - 200 seconds on a machine with 16 GB RAM and 2x 4.20 GHz. Repeating an analysis returns the same result except for differences at alignment edges within repetitive or duplicated regions, e.g. at contig edges.

There are some features which have to be considered when using the software. The best hit retrieval, depending on the complexity of the compared genomes, might fail at repetitive sequence regions. According to the nature of repetitive elements, such regions should be investigated in detail manually if of interest. Furthermore, the two sequences to be compared have to have a similar or identical genomic arrangement, since only left or right neighboured sequence regions are compared in an iterative manner. Rearrangements lead to not aligned regions and thus unresolved divergent sequence regions. This was observed in the analysis of the ST226 sequences, were an apparent mis-assembly occurred in one sequence. The sequences at the affected locations could not be aligned and led to two large unaligned regions. Subsequent comparison of the correct (manually assigned) sequence pairs at the two locations led to the correct results, which were confirmed by additional sequence information (unpublished) of 454 reads (see chapter 3.2). Also, batch execution or comparison of more than two sequences is not yet possible. Pairwise analysis and visualization including total

overview visualization of all input sequences (>2) might be an interesting feature for further development.

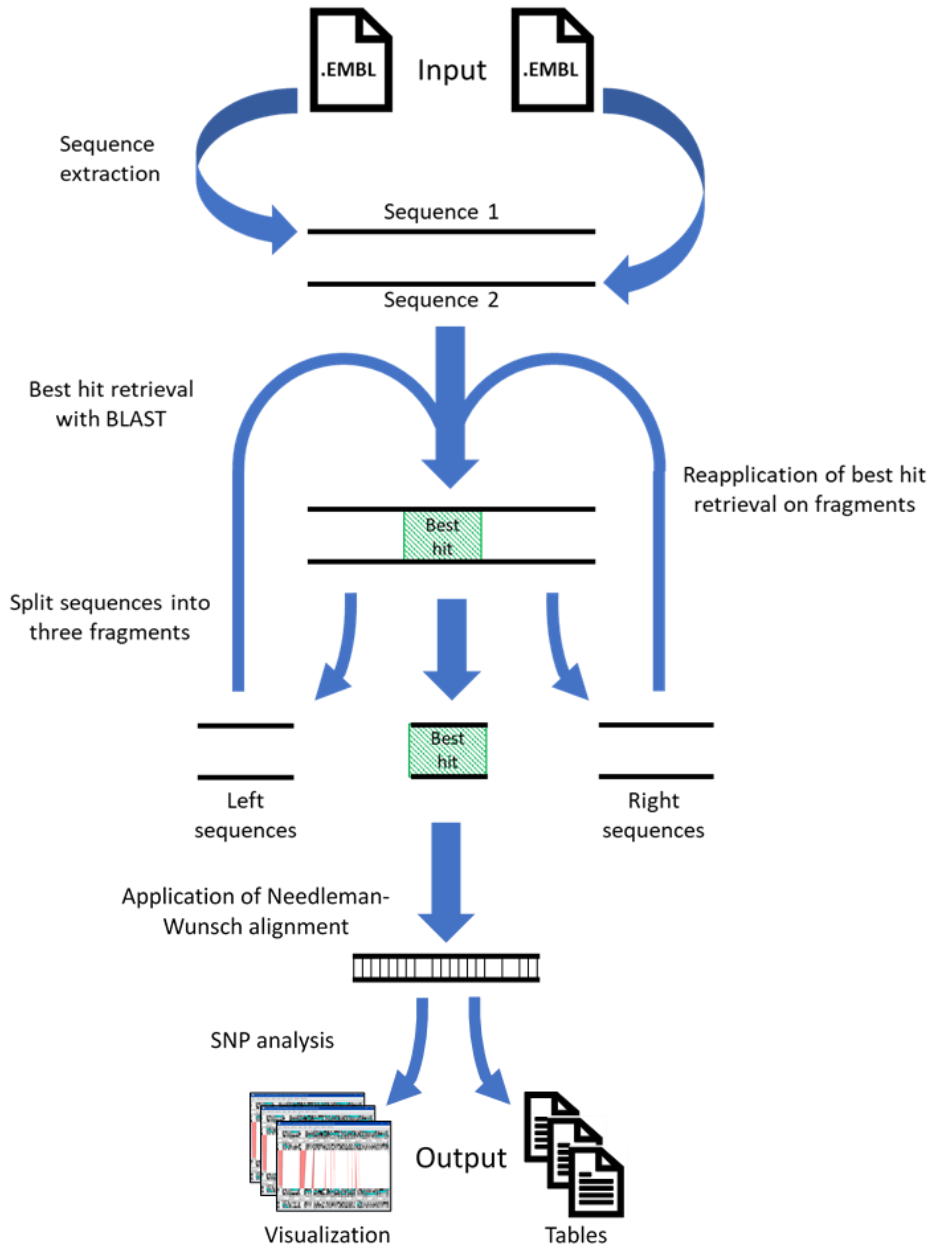


Figure 3.16: Model of the SNP analysis procedure

The two files in EMBL format, which serve as input of the analysis software and provide the sequences to be compared, are initially parsed to extract the nucleotide sequences and, if possible, the annotated features. Then, the best matching regions of these two sequences (best hit) were determined by the program BLASTN. If any matching regions are found, the sequences were split into three fragments: left and right of the matching region and the matching region itself. The left and right fragments serve as input of further best hit retrievals, while a Needleman-Wunsch alignment retrieves SNPs and indels from the best hit region. This procedure loops through the whole sequence pair until no left or right fragments are available anymore. The results of the SNP retrieval are collected together with information about unaligned regions and split into several types of output files. These files contain the results in tabular form but also visualization files of SNPs and unaligned regions for usage with the Artemis Comparison Tool (ACT, Sanger Institute) are generated.

4 Discussion

4.1 Sequencing, assembly and annotation

Through the last decades, several techniques have been developed based on diverse methods to retrieve sequence information of DNA, RNA and proteins as described in chapter 1.2.1. This development facilitates a rapidly increasing number of sequences with decreasing cost at the same time (Huse, et al., 2007; Salzberg, et al., 2012). Third-generation sequencing technologies will lead to further acceleration of this process (see introduction). Here, the estimation of genome sequences is discussed, concerning expected and generated results, problems and error susceptibility as well as solutions and further development.

The assemblies and analyses described in this work used data generated by 454-pyrosequencing and Illumina technology. Besides problems like sample quality or sequence structure (repeats etc.), each sequencing technology brings along its own problems (Huse, et al., 2007; Dohm, et al., 2008; Miller, et al., 2010) as described in the introduction. Both techniques generate sequences with an average length between 330 - 800 nt (Miller, et al., 2010; Luo, et al., 2012; Metzker, 2010) compared to first generation techniques with 500 – 1.000 nt (Miller, et al., 2010; Luo, et al., 2012). Reads generated by the 454-technology are longer than Illumina-generated reads and therefore are more capable of spanning repetitive sequence sections (see chapter 1.2.1), but the higher error rate especially in homopolymer stretches leads to indels and consequently to wrong annotation (see chapter 1.2.3). Due to the underlying technology, Illumina generated reads are less susceptible to such errors, although there is a common basic error rate. A high coverage helps to compensate this to some degree and can be reached especially in case of Illumina sequencing (Dohm, et al., 2008; Miller, et al., 2010). Other methods like read trimming to reduce carry forward errors and incomplete extensions also are applied for error detection and correction and are integrated in sequencers and assemblers to some extent (Dohm, et al., 2008; Salzberg, et al., 2012). Standalone applications can improve this integrated error-detection.

The 454-generated reads used in the publications (Rieger, et al., 2017; Denpaite, et al., 2016) (see chapters 3.1 and 3.3 and supplementary table S1) have lengths of only between 182 and 244 nt, which is much less than the general average length of 330 - 800 nt of 454 reads (Miller,

et al., 2010; Luo, et al., 2012; Metzker, 2010), while the Illumina-generated reads have an average length of 151 nt (Todorova, et al., 2015; Rieger, et al., 2017) (see chapters 3.2 and 3.4 and supplementary table S1) as expected for this technology (Metzker, 2010; Tritt, et al., 2012). After initial trimming by the assembler (Newbler 2.6), the average lengths of 454 reads decreased to 133 – 216 nt, while the Illumina read length decreased to 141 – 144 nt (supplementary table S1). This means that 11 – 16% of the nucleotides of reads generated by 454-technology and 4 – 6% of the nucleotides of the reads generated by Illumina-technology were not used in further analyses (supplementary table S1). The trimming of reads is based on the quality score where a high score indicates a lower error probability (Huse, et al., 2007; Dohm, et al., 2008). In 454-generated reads, the quality score indicates the probability of a correct length of homopolymer stretches, whereas it represents the probability of a correct base call in Illumina-generated reads (Huse, et al., 2007; Dohm, et al., 2008). Our data confirm that Illumina technology is less error-prone when compared to the 454-technology which is rarely used any more.

The assembled genomes presented in the current work also show a remarkable difference of the coverage depending on the sequencing technology as expected (Luo, et al., 2012; Chaisson, et al., 2008; Liu, et al., 2012), where coverage in the current case is the ratio of nucleotides of aligned reads and of nucleotides of generated contigs due to unknown length of the target genomes. The average coverage (the depth of coverage is not uniformly distributed over the target sequence) usually is determined by the ratio of nucleotides within reads and the known target genome size, where the coverage depth depends on the accuracy of the prior assembly (Sims, et al., 2014). The Illumina generated sequences (PCP and Hu) show with 144 – 169x a ten-fold higher coverage than the 454-generated sequences (D-Isolates, primate streptococci) with about 7 – 18x (see supplementary table S7 and chapters 3.1, 3.2 and 3.4). It should be noted that the coverage rate is only an approximation. Although complete genomes are available for *S. pneumoniae* (R6 and TIGR4; (Hoskins, et al., 2001; Tettelin, et al., 2001)) as well as for *S. mitis* and *S. oralis* (Denapaite, et al., 2010; Reichmann, et al., 2011), genomes of other strains may vary in size due to a highly variable accessory genome in these species. Minimum coverage values recommended for an assembly are between 15 - 60x (Ajay, et al., 2011; Kisand, et al., 2013; Fang, et al., 2014; Bentley, et al., 2008), which is in the range achieved in the work presented here. Moreover, as stated by Ajay

et al. (Ajay, et al., 2011), the completeness of an assembly is not as important for analyses on a population level as for determination of individual genomes.

Especially indels leading to stop codons and wrong positions and falsely annotated start codons of protein coding genes aggravate comparative analyses between genomes and require manual inspection if individual genes are being analysed. For example, this can be seen at the comparison of the ST10523-isolates D122 and D141 (see supplementary table S2). This comparison revealed 46 SNPs and 39 indels in 63/56 genes. While most of the SNPs affect only one or even none encoded amino acid, every indel changes the subsequent gene sequence and thus the deduced protein, and two lead to stop codons. Another problem during the assembly are repetitive sequences within genes or genes that occur more than once within the genome, resulting in gaps. Examples are RNA clusters and *comX* of *S. pneumoniae*, and the repetitive motifs present in e.g. CBPs or within *S. pneumoniae nanA* frequently lead to wrong assembly and consequently to wrong annotation. During assembly, the generated sequence graph might lead to diverging branches (see introduction), after which the assembly breaks and contigs are generated. Genes located at such sequence breaks (contig edges, gaps) cannot be correctly or not at all be identified by annotation algorithms. Manual search or synchronization with correctly annotated sequences might compensate this deficiency. In case of the three ST10523 sequences, the automatically generated annotation by *RAST* (Aziz, et al., 2008) had to be reviewed manually for synchronizing divergent or missing gene annotations prior to a detailed comparison of individual genes. The annotations of the streptococci from primates were not synchronized to each other due to large manual effort and thus differed frequently from that of known streptococcal genomes. Unfortunately, the annotation of genomes publicly available is not synchronized with the guidelines recommended by the upload platforms e.g. of NCBI. Therefore, in the comparative analysis of individual genes the annotation of the finished *S. pneumoniae* R6 (Hoskins, et al., 2001; Lanie, et al., 2007), *S. mitis* B6 (Denapaite, et al., 2010) and *S. oralis* Uo5 (Reichmann, et al., 2011) genomes was used as standard. Here, the call of Kisand *et al.* (Kisand, et al., 2013) for "standardization of gene prediction and annotation" can be extended to the synchronization of gene prediction and upload platforms. Complete and well curated genomes might contribute to the quality of public available annotation databases and tools (Kisand, et al., 2013). The number of available incomplete genomes will certainly grow faster than that of

complete genomes, leading to steadily increasing confusion on the level of gene annotation. Contributing to this, scientists might push forward annotation of "their" genomes only until the desired information has been obtained (Kisand, et al., 2013). Due to the high effort to complete genome sequences (e.g. by manual sequencing of gaps) and add and adjust the annotation of their features, one has to live with this problem at the moment. But this does not mean that the curation cannot be done for particular analyses and genomes. On the other hand, the PCP-transformants (Todorova, et al., 2015) and the ST226 (Rieger, et al., 2017) sequences were annotated by transfer using the annotation of *S. pneumoniae* Hungary^{19A}-6 (NC_010380), *S. pneumoniae* R6 (Hoskins, et al., 2001; Lanie, et al., 2007) and *S. oralis* Uo5 (Reichmann, et al., 2011). In the first case, the annotation was performed manually aided by *BLASTN*, in the second case by *RATT* (Otto, et al., 2011) with subsequent manual curation. An apparent disadvantage of this procedure is that wrongly annotated genes of the reference genome remain incorrect.

In summary, although *Newbler* is be able to handle Illumina data as well as 454 data, different problems occur depending on the technology used. While the error rate of 454 data is remarkably high especially at homopolymer stretches, Illumina data are not always able to span even short repeats. Independent on the sequencing technology, a major problem of composing and finalizing genome sequences is the annotation, which has to be automated and curated much more and carefully than currently done. Furthermore, genes with dubious sequences were not used for further analysis. Concerning 454-generated sequence data, this includes proximity to a sequence gap or differences in homopolymer stretches. In general, repetitive or mobile elements also aggravate analysis and were thus excluded. But as demonstrated, filtering the data leads to a loss of information which has not to be underestimated (6 – 16% of nucleotides in the current work). Direct sequencing of the excluded sequences might reduce this number as well as the choice of a more suitable sequencing technology (excluded nucleotides: 454: 11 – 16%; Illumina: 4 – 6%). But dependent on the particular focus of research, these values might be acceptable after reviewing the excluded sequence. The appearance of third-generation sequencing technologies might solve some of the mentioned problems. Increased read lengths (several thousand nucleotides) (Land, et al., 2015; Lu, et al., 2016) might facilitate assemblers to resolve repeats and other ambiguities and generate bacterial genome sequences without or

with few gaps. But still, the error-per-base rate seems quite high as well as the costs compared to short-read technologies like Illumina and thus the usage of third-generation sequencing technologies has to be well-considered (Boldogkői, et al., 2019).

4.2 Analyses

During the last years, the number of sequenced genome sequences steadily and rapidly increased and thus comparative genomics has to deal with huge data sets. Subsequent analysis might run into problems to handle these data due to different generation and analysis methods, tools and formats, which complicate comparison of the increasing amount of data. Furthermore, more data do not imply better quality or more and accessible results. The focus should be on maintaining high and increasing quality and reliability of curated data than on masses of unfiltered raw data.

Often the required information like reads, qualities, coverages or flowgrams for analyses or further error correction are not available (e.g. comparison with 23F capsule in chapter 3.1.3 or pilus in chapter 3.3.2). Concerning the majority of analyses described in this document, workflows were defined, which tolerate a certain loss and uncertainty of data, but provide reproducibility and comparability within a certain scope. There is a variety of open source tools available providing functionality for several analysis purposes and, besides decreasing sequencing costs and increasing quality (Kisand, et al., 2013), enable even smaller laboratories and single researchers to perform analyses. But no one met all requirements described in chapter 3.6 and thus a new program was developed, employing an approved mechanism of recursive alignments connected with conflation of annotation data and sequence comparison at low level resulting in desired and visualizable output format. The program can be started at command line and with minimal configuration effort. Besides the input files and output path, only the threshold for the distance of a detected difference to a gap has to be defined. Therefore, individual adjustment of alignment parameters is not possible, what makes the analysis results comparable. The fast and uncomplicated workflow of the software generates easy to read result tables and visualization, where differences are displayed between the compared sequences. Furthermore, the input sequences, as long as they are provided in EMBL format, are not restricted to a specific sequencing technology, but the program is not able to resolve rearrangements within the sequences. These are only found as not matching regions and have to be resolved manually. The disadvantage of using sequences decoupled from sequence reads and quality information is the general use or non-use of sequences, which might lead to detection of false positive or negative results. The steadily increasing number of

public available and annotated genomes makes it easier to find genomes with similar sequence organization for comparison and further analysis, if needed.

Depending on the relatedness between members of bacterial clones, strains and species, the number and extent of differences of compared genome sequences naturally increases varies. Thus, one expects that genomes of the same clone differ by only few SNVs and no large-scale difference, and that between genomes of different strains the number of SNVs will be increased while also differences of larger sequence regions can be expected e.g. by horizontal gene transfer etc. A special case are transformants produced in the laboratory which might differ by a few SNVs besides transferred regions.

S. pneumoniae ST10523 (Rieger, et al., 2017) and ST226 (Rieger, et al., 2017) represent comparisons of genomes of the same clone. As expected, the comparisons of members of the two clones revealed only few SNVs within genes (ST10523: 85 (D122/D141), 235 (all three genomes); ST226: 37). Unfortunately, the results of the analyses of the *Streptococcus pneumoniae* clones ST10523 (Rieger, et al., 2017) and ST226 (Rieger, et al., 2017) cannot be compared unrestrictedly. At the one hand, the genomes of the ST226 clonal complex generally are more variable as of other clones (Hakenbeck, et al., 2001). Moreover, the analysed genomes of the two clones are based on different sequencing technologies (ST10523: 454; ST226: Illumina). As already described (previous chapter), the differences between 454-generated sequences which are located in homopolymer stretches, had to be removed from analysis due to uncertainty, as well as differences located near sequence gaps. Additionally, due to the high error rate of 454-generated sequences, the results of searches for SNVs (single nucleotide variations; polymorphisms and indels) might be generally questioned (Kisand, et al., 2013). At another part of the current work (see chapters 2.3 and 3.3), the first attempt to find and compare sequences of the autolysin encoding gene *lytA* in streptococci from primates (Denpaite, et al., 2016) was extremely complicated due to sequence gaps and differing homopolymer stretches. In the current work mainly SNV within genes were further analysed since changes in the encoded protein were the main issue. Naturally, differences in intergenic regions affecting the transcription are of importance but were not further considered in the current analyses. There are several cases of differences between sequences, which were analysed manually afterwards. For example, the *hlyA* gene of ST10523 was excluded from the global analysis but investigated manually due to the importance of the gene product as a main

virulence factor afterwards. Otherwise, important information would be missing. Concerning the same gene, intergenic differences between the compared genomes are located at the promoter region affecting expression of the *hlyA* gene. Besides SNVs, short additional sequences could be found present in one and absent in another genome during comparison of ST10523 and ST226.These sequences mainly occur in repetitive regions or at gaps. Since further effort is necessary to verify these differences and to exclude possible sequencing errors, they were not analysed. Treangen *et al.* (Treangen, et al., 2011) stated, that just ignoring repetitive sequence regions might distort analysis results and thus would be no option. They suggest resolving this problem e.g. by using paired-end information or multiple sequencing technologies. The ST226 genomes were generated from Illumina reads with paired-end information and this might be an explanation, why only 130 of these additional sequences were found (242 – 415 at the 454-generated ST10523 genomes), although especially the short Illumina reads are known to have problems spanning longer repeats. Leaving aside gapped and repetitive regions, the ST226 sequences differed in ten additional sequences from each other, while the ST10523 sequences differed only in 2 – 5 sequence regions. This is not due to the underlying sequencing technology. As mentioned above, the ST226 clone shows a high variability in its genomes which was confirmed in this analysis.

Another and special case of SNV retrieval is the analysis of *S. pneumoniae* PCP-transformants (Todorova, et al., 2015). While for analyses of ST10523- and ST226-sequences the program described in chapter 3.6 was used, the PCP-sequences were compared using *BLASTN* with subsequent manual determination of differences including sites of recombination. The first comparison of each genome was made with the genome sequence of the recipient *S. pneumoniae* R6, where SNVs were identified. During the subsequent comparison of the sequence fragments which could not be aligned with *S. pneumoniae* R6, with the genome sequence of the donor *S. oralis* Uo5, transferred sequence regions were identified. Afterwards, the three transformants were compared pairwise to each other and differences determined. In contrast to the analyses of the clones as described above, only few differences of the transformants to each other, to the donor and the recipient genome were expected and thus this alignment strategy seemed sufficient.

The analyses of the primate streptococci (Denpaite, et al., 2016) were complicated by presence of highly diverse *S. mitis* sub-clusters, which aggravate strain identification, as well

as by 454-specific problems, which make annotation and analysis difficult. Except for the pilus-2 islet and particular genes, the focus of this work was not on SNV level but rather on presence or absence of genes at the level of strains and species (in opposite to strains belonging to one clone as described before). The comparison of *S. pneumoniae* R6 and its close relatives *S. mitis* B6, *S. oralis* Uo5 and *S. pseudopneumoniae* IS7493 and finished *S. pneumoniae* genomes (Tettelin, et al., 2015) also operates on strain/species level. In contrast to the analyses of the clones or transformants, the focus was on presence or absence of genes rather than differences on the SNV level as described in the next chapter. Thus, the gene content of the genomes was determined including a certain tolerance regarding sequence variation. A tolerance of 60% coverage and 70% identity at pairwise protein alignment proved advantageous to define proteins of a reference genome as present or absent (Denapaite, et al., 2010) . Proteins present in all compared strains constitute the core-genome, and proteins present only the reference genome are considered as strain specific and could be dispensable.

4.3 Genomic diversity

The current work focuses on comparisons of genomes, their encoding genes and deduced proteins. This task was facing several problems. Some of them concern sequence and annotation problems already described in the chapters 4.1 and 4.2. Another point is the purpose of the comparisons, which depended on the particular genomes to be analysed as outlined in the subsequent chapters.

4.3.1 Technical issues

A major problem of comparing several genomes is their state of completeness. Incomplete (gapped) genomes contain regions without known sequence information, usually represented by a certain number of 'N'. Even with similar genomes, these regions are not always located at the same position and thus a sequence is present in one and absent in another genome. Furthermore, underlying sequencing technologies lead to additional problems. Concerning 454 technology, homopolymer stretches which are not validated are not reliable, as well as short reads of the Illumina technology which lead to an increased number of gaps (as described in chapter 4.1). Stringency of comparison parameters and consideration of variability of genes and proteins are closely linked and the choice of the basis for a suitable comparison (proteins, genes, genome sequence in general, etc.) is also important. If annotated features (like CDS or genes) are used, different annotation of the same feature in different genomes is a further and not to be underestimated problem. Finally, distinct sets of genes/proteins need to be considered depending on the particular analysis.

Whereas complete genomes were used in chapters 3.4 and 3.5 and (Tettelin, et al., 2015; Todorova, et al., 2015), comparisons in other cases were based on incomplete genomes and datasets in the first instance. The problem with incomplete sequences differing in the locations of known and unknown sequence regions was consistently solved in several steps. Comparisons were focused on presence or absence of genes or proteins (features) and differences between them. Features were manually reviewed, annotation differing between compared genomes was adjusted, and features were excluded from the analysis if they were completely or partially located within a sequence gap in one of the analysed genomes (see chapters 4.1 and 4.2). In case of genomes from oral streptococci from primates, only mobile

elements were removed from CDS. This resulted in a loss of a certain amount of information per genome (see chapter 4.1) but was a useful step for the comparisons. Since therefore not all features were used for analysis, the ratio of features used for analysis and features containing differences between genomes is much more expressive than their absolute number. The comparisons are restricted so sequences and features, which are not filtered in any sequence. 454 and Illumina data are not mixed within one analysis.

The decision whether a protein was present or absent in a particular genome, a similarity of 60% coverage (concerning the sequence length) and an identity of 70% (similarity of sequence content) has been proven to be a good choice (Denapaite, et al., 2010). These values allow a certain variability of the analysed proteins. However, a problem occurred when an indel was present causing a frameshift and thus a stop codon within an encoding gene resulting in the annotation of only part of the protein. Consequently, if the length of the protein fragment was smaller than 60% of the complete proteins it was missed in the comparison. In order to solve this problem manual inspection was required to verify absence/presence on the DNA level.

4.3.2 Analysis of *Streptococcus pneumoniae* R6 Transformants obtained with DNA of completely known genome sequences

The genomes of *S. pneumoniae* R6 and *S. oralis* Uo5 are complete and annotated (Hoskins, et al., 2001; Lanie, et al., 2007; Reichmann, et al., 2011). Transformation experiments using donor DNA from the high level penicillin resistant *S. oralis* Uo5 and the sensitive laboratory strain *S. pneumoniae* R6 as recipient revealed transfer of several genomic regions after three transformation steps (Todorova, et al., 2015). Not only the genes encoding Pbp2x, Pbp2b, and Pbp1a, but surprisingly MurE contributed to penicillin-resistance. This finding shows a yet unrecognized resistance determinant in *S. pneumoniae*, and one can expect that there might be further proteins contributing to penicillin resistance which are yet unknown. MurE was described earlier to contribute to β-lactam resistance in *S. aureus* (Gardete, et al., 2004). Interestingly, the *murE* gene alone as well as its promoter region alone are capable to increase penicillin-resistance; however, this effect is not cumulative. Besides these four genes, many genes or parts of genes were transferred, which were not associated with penicillin-resistance. This was also observed with *S. pneumonia*e R6 transformants obtained with *S. mitis* B6 DNA (Sauerbier, et al., 2012). However, while the transformation with *S. mitis* DNA led to the recombination of about 66 kb in 16 clustered regions (containing closely located recombination events) ranging between 160 bp to nearly 23 kb, the transformation of *S. oralis* DNA led to the transfer of only approximately 19 kb in 9 regions with sizes of 104 – 3.322 bp (Meiers, 2015). This is not unexpected since *S. mitis* is more closely related to *S. pneumoniae* than *S. oralis* and thus the higher sequences similarity of *S. mitis* allows for more recombination events.

After three transformations steps to obtain the transformant PCP, another three transformations and selection with beta-lactams resulted in only a few 'transformants'. The genome sequence of the last transformant PCP-CCO revealed that it contained no further *S. oralis* Uo5 DNA, rather mutations in four genes had occurred during the selection procedure, and these mutations were confirmed by manual resequencing. PCP-CCO, contains one single point mutation each in *ciaH* and *pbp2b*, known to contribute to penicillin resistance, and in *spr1992* encoding a protein of unknown function that apparently in combination with *ciaH* also contributes to resistance (Meiers, 2015). Furthermore, a deletion occurred in *htrA* encoding the serine protease HtrA, leading to a frameshift and thus to a premature stop codon

125, most likely resulting in a non-functional product. The serine protease HtrA has been described as a virulence factor (Ibrahim, et al., 2004), because HtrA mutants have a decreased ability for colonization (Sebert, et al., 2002).

4.3.3 Common genes of *S. pneumoniae* and close relatives

The identification of virulence factors (VF) in *S. pneumoniae* has been in the focus of research for many decades. Over one hundred virulence factors have been described in *S. pneumoniae* (Mitchell, et al., 2010; Brown, et al., 2002; Hava, et al., 2002; Polissi, et al., 1998; Lau, et al., 2001), mainly identified by a decreased pathogenicity potential using mouse models. Denapaite *et al.* described the presence of many of these VFs in the genome of the commensal organism *S. mitis* B6 (Denapaite, et al., 2010), a finding later confirmed by Kilian *et al.*, who extended the comparison to further *S. pneumoniae* isolates and several members of the closely related species *S. mitis*, *S. oralis*, *S. infantis* and *S. pseudopneumoniae* (Kilian, et al., 2019). Thus, only a few *S. pneumoniae* specific virulence factors remain.

To investigate overall commonalities and differences between members of *S. pneumoniae* and representatives of closely related species (*S. oralis* Uo5, *S. mitis* B6 and *S. pseudopneumoniae* IS7493), several complete genomes were compared on the basis of their deduced protein sequences. The degeneracy of the genetic code leads to the same deduced protein despite different DNA sequence and the used method facilitates to tolerate this variability during analysis. Common genes (encoding the compared proteins) shared between the genomes are referred here as 'core' in contrast to their large accessory genome (Hakenbeck, et al., 2001; Tettelin, et al., 2015). As described in chapter 3.5, several combinations of genomes were used to calculate core genomes: *S. pneumoniae* and representatives of related species with and without *S. pseudopneumoniae*, only *S. pneumoniae* genomes and genomes of different

Table 4.1: Overview of core genomes within and between streptococcal species

The collection contains core genomes of members of the same clone, of the same species (intra-species) and of different streptococcal species (inter-species). Members of the same *S. pneumoniae* clone share the highest number of genes, while *S. oralis* strains isolated from different hosts share the fewest. It should be noted that the comparison between 12 *S. oralis* includes isolates of different hosts.

genomes	scope	core proteins
ST10523	clone	1.547
ST10523 + 6 *S. pneumoniae*	intra-species	1.146
26 *S. pneumoniae*	intra-species	1.285
S. pneumoniae, *S. mitis*, *S. oralis* (, *S. pseudopneumoniae*)	inter-species	1.140 (1.105)
12 S. oralis	intra-species	823
26 S. pneumoniae, *S. mitis*, *S. oralis*, *S. pseudopneumoniae*	inter-species	966

species together with 26 *S. pneumoniae* genomes. An overview is listed in Table 4.1. It should be kept in mind, that the numbers are based on different genomes, sequencing technologies and filtering, leading to slightly different numbers of total and core genes. Adding genomes of the same species and of distinct clones to the calculation, the number of core-genes drops remarkably (74% of the clone-specific core). Using the draft sequences of three members of the *S. pneumoniae* clone ST10523, 1.547 common genes were identified. Adding only *S. pneumoniae* R6, the number of common genes drops to 1.323, and to 1.146 adding another five *S. pneumoniae* genomes. Using another set of 26 *S. pneumoniae* genomes derived from distinct clones excluding ST10523, 1.285 common genes were identified. An explanation that this number is higher probably reflects a more stringent use of the ST10523 sequences and the annotation of the ST10523 genomes with RAST. Concerning the interspecies core based on complete genome sequences representing four streptococcal species, 1.140 common genes were noted, and again this number dropped to 966 by including another 25 *S. pneumoniae* genomes. This analysis is especially important when analysing pneumococcal specific virulence factors.

The decrease of core genes upon addition of more genomes and other species was also described by Kilian *et al.* (Kilian, et al., 2019). The core of 60 genomes of the species *S. pneumoniae*, *S. mitis*, *S. oralis*, *S. pseudopneumoniae* and *S. infantis* is represented by 690 genes. Excluding *S. infantis*, the core genome of the remaining 54 genomes contains 894 genes, slightly lower than the value 966 calculated in the current work using 26 genomes of *S. pneumoniae* and only one representative genome of each related species reflecting the extended set of genomes of non-pneumococcal species in the study by Kilian *et al.*.

Another approach to estimate the pneumococcal core-genome with another method (Bayesian) was compared to a COG-based (COG: Cluster of orthologous groups) method (van Tonder, et al., 2014). The underlying dataset contained 616 pneumococcal genomes using *S. pneumoniae* ATCC700669 as reference (Croucher, et al., 2013; van Tonder, et al., 2014). The numbers of core genes (Bayesian: 948; COG: 1.194) differed noticeable due to different stringencies used for the definition for presence of a gene/protein. The COG-based core tolerates a certain variability of genes, while the stringency of the Bayesian-approach at first was 100% coverage and identity. After allowing a certain variability (90% coverage), the Bayesian approach resulted in 1.206 core genes, a value similar to the COG-approach. Despite

the similarity of total core gene numbers, the content of the two core gene sets differs by up to 179 genes unique to each estimation method. These numbers are lower compared to our data (1.285), probably due to the higher number of genomes analysed by these authors (Croucher, et al., 2013; van Tonder, et al., 2014). It should also be kept in mind, that the genes/proteins of the current work were extracted from a reference sequence and used directly to estimate the core genome without grouping into clusters. Nevertheless, the calculated number of core genes lies in a similar range around 60 % of the entire genome, independent on the method used. As stated by van Tonder *et al.* (van Tonder, et al., 2014), estimations of the core genomes depend on the data set and the parameters used and thus it is impossible to define a single and universal core genome.

The pneumococcus specific core as defined in the current work includes only a few of the described virulence factors: the CBPs PcpA, PspA, and PspC (and its variant Hic) together with the two-component system TCS06, the ply-lytA island, and the hyaluronidase. Moreover, the polysaccharide capsule is required for pathogenicity. Some of them such as the CBPs, PBPs, and MurMN are highly variable due to an apparent mosaic structure. This indicates frequent horizontal gene transfer events, one more justification for comparative genomics. MurN was missing in most *S. oralis* and one *S. mitis* genome, and a variant of the Ply-LytA island was found only in two *S. mitis* genomes. The TCS06 was present in all genomes but they were missing PspC, indicating unknown regulatory functions of this system. One gene of the polysaccharide capsule (*cpsO*) was found in *S. oralis* Uo5 (see supplementary table S11). Recent data show that a *cps* cluster is present in several streptococcal species, suggesting that it has been imported into *S. pneumoniae* from other sources (Skov Sørensen, et al., 2016; Kilian, et al., 2014). In rare cases unencapsulated *S. pneumoniae* isolates are able to develop a certain pathogenicity potential (Keller, et al., 2016). The presence of most virulence factors also in commensal species like *S. mitis* emphasizes their role in colonization and interaction with host tissue.

The current work confirms the evolutionary model of a common ancestor of *S. pneumoniae* and *S. mitis* (Kilian, et al., 2014). This ancestor, putatively pathogenic to the human ancestor, was separated into two lineages as response to selective pressure. The *S. mitis* lineage became a commensal organism coexisting with the human host and partially loosing genes associated with pathogenicity and virulence. In contrast, the *S. pneumoniae* lineage obtained the

potential to be more virulent and pathogenic by extending its potential of horizontal gene transfer within and between related species.

4.3.4 New genomes of two particular clones

Concerning the high degree of relationship between strains belonging to the same clone (ST10523 and ST226), a detailed comparison on the DNA level to detect SNVs was required to reveal their differences as described in chapter 4.2. Besides manual analysis (e.g. of previously excluded features), these comparisons revealed a relatively low variability between the genomes as expected but also interesting differences.

The main question concerning the ST10523 clone was, if special genes and mutations within this clone could explain the unusual long persistence within the same host of two strains (chapters 2.1 and 3.1). The analysis was based on the strategy for removing potential error-prone data as described in chapter 3.1.1.2. Only a five-gene-cluster was missing in one isolate besides the presence of two phage clusters, which were excluded from analysis. Phages and mobile elements constitute a considerable proportion of the accessory genome of a clone (Croucher, et al., 2011). We calculated that 92% (1.547 CDS) of the clonal pan-genome is present in all three genomes of clone ST10523. 56 – 63 genes were affected by 85 SNVs (comparison of D122 and D141) and 153 – 163 genes by 235 SNVs, respectively (all three isolates). None of these genes appear to contribute to the long persistence of the ST10523 isolates. Manual inspection of genes possibly involved in virulence revealed an unusual hyaluronidase gene, one of the *S. pneumoniae* specific virulence factors. It contains deletions in the promoter region (12 bp) as well as a deletion of 4 nt in the coding region, most likely leading to a non-functional protein. During the initial analysis, the hyaluronidase gene was not considered due to a gap in the genome sequence of *S. pneumoniae* D219. All three ST10523 genomes contained the same allele. Furthermore, the ST10523 isolates carry a unique variant of the surface protein PspA. It is possible that the absence of two phage clusters also contributes to the long-term survival of the two ST10523 isolates D122 and D141.

The second clone analysed in detail, ST226, is high-level penicillin resistant but contains interestingly one member which is susceptible to beta-lactams (chapters 2.2 and 3.2). Therefore, the task was to see, which genes are responsible for the phenotypic divergence. Genomes of the Hungary19A clonal complex which includes the SLV ST226 show a much higher variability compared to other *S. pneumoniae* clones (Rieger, et al., 2017; Hakenbeck, et al., 2001). Comparing the genomes of Hungary^{19A}6, and of the two ST226 strains Hu15 (penicillin-sensitive) and Hu17 (high-level penicillin resistant), 71 and 75 CDS were inserted,

deleted or exchanged, a much higher number compared to the ST10523 clone. The two ST226 genomes Hu15/Hu17 differed in a gene encoding a putative N-acetyl-neuraminidase which revealed high SNV density. The role of pneumococcal neuraminidases in colonization and contribution to otitis media could be ascertained in chinchilla model (Tong, et al., 2000). Both genes match homologues with almost 100% identity and coverage when compared to the NCBI *S. pneumoniae* data base, suggest horizontal gene transfer events in this region. Schweizer *et al*. (Schweizer, et al., 2017) analysed the clone ST226 based on the data presented in the current work. PBPs, the main penicillin-resistance determinants, have a mosaic structure in *S. pneumoniae* Hu17 but not in *S. pneumoniae* Hu15. However, both strains contain a unique allele of CiaH (named CiaH232) and a MurM variant, suggesting their presence prior to the introduction of mosaic PBPs in this clone. Comparison of MurM of ST226 with genomes of other streptococci revealed that it most likely originated from *S. mitis*. The presence of these two genes did not contribute to penicillin resistance in the absence of mosaic PBPs but was required to guarantee proper cell morphology. CiaH232 has been shown to affect CiaR-mediated transcription, suggesting that that the *cia*-system is somehow involved in the regulation of cell wall synthesis. Surprisingly the resistance level increased substantially when MurM of ST226 was combined with a mosaic PBP2x, and CiaH232 could contribute to resistance when introduced into a strain carrying both, *pbp2x* and *pbp1a* from Hu17. Thus, based on the genomic comparison, the identification of genes involved in penicillin resistance and further genetic experiments revealed novel aspects on the origin and regulation of this phenotype.

4.3.5 Genomes of Streptococci of different hosts

The analysis of streptococcus isolates from human versus primate hosts which included wild animals that had no contact to humans revealed several new aspects of the evolution of Streptococci and *S. pneumoniae* virulence factors. The initial speciation analysis by MLSA and MLST revealed that many viridans streptococci that are common in human were found in great apes and monkeys, but not in lemurs from Madagascar. Several isolates could not be specified and clustered outside of known streptococci. *S. oralis* was common among wild chimpanzees and other monkeys, indicating that this species has evolved before the appearance of humans. Interestingly, some of the primate *S. oralis* formed three clearly separated groups challenging the definition of this species, which is much more diverse than *S. pneumoniae.* In contrast, *S. mitis* isolates were only obtained from one gorilla held in captivity, suggesting that *S. mitis* and *S. pneumoniae* have evolved in humans. The genomic analysis focused on virulence factors as defined in *S. pneumoniae* and on components involved in cell surface components.

The choline binding proteins PspA, PspC and PcpA were only found in *S. pneumoniae*, confirming species specificity. Related proteins rarely occurred in other species.

The pneumolysin and LytA island was only found in two human *S. mitis* strains confirming former observations that single *S. mitis* strains contain these genes (Whatmore, et al., 2000; Neeleman, et al., 2004; Tettelin, et al., 2015; Kilian, et al., 2008). Moreover, the pneumococcal hyaluronidase HysA (HlyA) was present only in one *S. oralis* strain. Interestingly, only *S. pneumoniae* and *S. mitis* harboured the N-acetyl-neuraminidase gene NanBC whereas *S. oralis* contained a gene encoding another β-N-acetyl-hexosaminidase Pili can contribute to virulence by facilitating adhesion to host tissues. A new variant of the pilus islet 2 has described recently, and variants of the major pilus subunit PitB were found in several primate streptococci. Part of the cia-regulon in *S. pneumoniae* are five so-called cia-dependent non-coding small RNAs (ncRNA/csRNA), which also contribute to virulence potential (Marx, et al., 2010). The current work revealed their widespread occurrence and variants thereof in viridans streptococci. One surprising result was the presence of six csRNAs due to duplication, and genetic islands were integrated between the duplicated csRNAs, suggesting that these structures function as entry site during horizontal gene transfer.

Since *S. oralis* could be isolated from several primates, it was interesting to see, whether, and if how, they differ from the human isolate *S. oralis* Uo5.

Genes required for choline-containing teichoic acid described in *S. pneumoniae* were common in *S. mitis*, *S. oralis* and present in even one *S. infantis*, and in case of the *lic4* gene cluster these genomes always contained the three CBPs LytB, CbpD and CbpF. Astounding was the high variability of proteins involved in peptidoglycan-synthesis (PBP1a, 2b, 2x; MurMN) especially in *S. oralis*, previously described for *S. mitis* (Denapaite, et al., 2010; Kilian, et al., 2014). The human isolate *S. oralis* Uo5 contains an "unusual" *murM* gene and no *murN* gene (Reichmann, et al., 2011), and in some primate *S. oralis murM* and *murN* are missing. It has been suggested that the absence of MurM is only tolerated in a penicillin-sensitive context, since deletion of *murM* results in a breakdown of the resistance phenotype (Filipe, et al., 2000; Weber, et al., 2000). Interestingly, PBP2x and PBP2b of some MurMN-lacking strains contain point mutations which are known to contribute to penicillin-resistance. Their effect on penicillin susceptibility peptidoglycan structure needs to be investigated experimentally. In this context it should be pointed out the antibiotic resistance determinants were only found in isolates from animals held in captivity but not in wild animals. In contrast to the three PBPs mentioned above, PBP2a was highly conserved among all isolates. Whereas one PBP3 gene is present in *S. pneumoniae*, *S. mitis* and *S. oralis*, primate isolates representing other streptococci of the Mitis group contained two PBP3 variants but only one variant appears to be functional. The genetic environment indicated that these variants have been introduced by horizontal gene transfer on several occasions.

In summary, several new insights into cell surface components, the distribution of virulence factors and the genomic architecture complicated by inter-species gene transfer events have been obtained. No gene indicating host specificity could be identified, requiring more isolates especially from free living animals. However, new features related to *S. oralis* and other viridans streptococci extended previous studies that focused on *S. pneumoniae* and *S. mitis*.

5 Future prospects

The improvement of sequencing technologies, error correction, analysis methods and annotation are an ongoing process. The genomes sequences presented in the current work were generated by NGS technologies. Therefore, there are still gaps in the sequences as well as potential errors in homopolymer stretches in case of 454 technology. Third generation technologies are capable of decreasing the number of gaps and errors. Furthermore, repeat structures, which often lead to gaps in the genome sequence due to alignment problems, could be sequenced when the technology provides longer reads. As soon as the high error rate and costs (Boldogkői, et al., 2019) will be improved, this technology might provide sequences without gaps and ambiguous homopolymer stretches. Also, due to the generation of single reads per genome, assemblers and the errors they bring along will be obsolete. Less errors will also improve comparative studies especially in case of deduced proteins, where indels based on false sequences previously led to erroneous length of the gene products. In the studies presented here, many genes and differences between sequences had to be excluded from the analyses due to possible errors or gaps (4 – 16% of nucleotides) – 207 -233 CDS of the ST10523 strains were excluded due to homopolymer stretches. As it can be seen at the *hlyA* gene of the ST10523 isolates, important information can be missed and requires manual inspection of individual genes of interest. Thus, eliminating the sources of errors and gaps would massively improve quantity and quality of analysis results. The workflow presented in the current work to provide a detailed comparison of genome sequences provides a basis that can be adjusted in case of new technologies. Meanwhile, thousands of *S. pneumoniae* genomes are available, providing the opportunity to obtain insights into more global aspects of evolutionary mechanisms and gene transfer events, and a more precise determination of species-specific and core genes of bacterial species as well as of closely related species. Such data concern fundamental questions addressed in the present work: The evolution of antibiotic resistance, pathogenicity and virulence factors, and the evolution of human specific bacterial species.

6 Abstract

The number of sequenced genomes increases rapidly due to the development of faster, better and new technologies. Thus, there is a great interest in automation, and standardization of the subsequent processing and analysis stages of the generated enormous amount of data. In the current work, genomes of clones, strains and species of Streptococcus were compared, which were sequenced, annotated and analysed with several technologies and methods. For sequencing, the 454- and Illumina-technology were used. The assembly of the genomes mainly was performed by the *gsAssembler* (*Newbler*) of Roche, the annotation was performed by the annotation pipeline *RAST*, the transfer tool *RATT* or manually. Concerning analysis, sets of deduced proteins of several genomes were compared to each other and common components, the so-called core-genome, of the used genomes of one or closely related species determined. Detailed comparative analysis was performed for the genomes of isolates of two clones to gather single nucleotide variants (SNV) within genes.

This work focusses on the pathogenic organism *Streptococcus pneumoniae*. This species is a paradigm for transformability, virulence and pathogenicity as well as resistance mechanisms against antibiotics. Its close relatives *S. mitis*, *S. pseudopneumoniae* and *S. orali*s have no pathogenicity potential as high as *S. pneumoniae* available and are thus of high interest to understand the evolution of *S. pneumoniae*. Strains of two *S. pneumoniae* clones were chosen. One is the ST10523 clone, which is associated with patients with cystic fibrosis and is characterized by long-term persistence. This clone is lacking an active hyaluronidase, which is one of the main virulence factors. The lack of two phage clusters possibly contributed to the long persistence in the human host. The clone ST226 shows a high penicillin resistance but interestingly one strain is sensitive against penicillin. Here it could be seen that the penicillin resistance mainly arose from the presence of mosaic-PBPs, while special alleles of MurM and CiaH - both genes are associated with penicillin-resistance – were present in resistant and sensitive strains as well. Penicillin resistance of *S. pneumoniae* is the result of horizontal gene transfer, where DNA of closely related species, mainly *S. mitis* or *S. oralis*, served as donor. The transfer of DNA from the high-level penicillin-resistant strain *S. oralis* Uo5 to the sensitive strain *S. pneumoniae* R6 was intentioned to reveal the amount of transferred DNA and whether it is possible to reach the high resistance level of *S. oralis* Uo5. Altogether, about 19kb of *S. oralis* DNA were transferred after three successive transformation steps, about 10-fold

less than during transfer from *S. mitis*, which is more closely related to *S. pneumoniae*, as donor. MurE was identified as new resistance determinant. Since the resistance level of the donor strain could not be reached, it is assumed, that further unknown factors are present which contribute to penicillin resistance. The comparison of *S. pneumoniae* and its close relatives was performed using deduced protein sequences. 1.041 homologous proteins are common to the four complete genomes of *S. pneumoniae* R6, *S. pseudopneumoniae* IS7493, *S. mitis* B6 and *S. oralis* Uo5. Most of the virulence and pathogenicity factors described for *S. pneumoniae* could also be found in commensal species. These observations were confirmed by further investigations by Kilian *et al.* (Kilian, et al., 2019). After adding 26 complete *S. pneumoniae* genomes to the analysis, only 104 gene products could be identified as specific for this species. Investigations of a larger number of related streptococci, which were isolated from human and several primates, confirmed the presence of most of the virulence factors of human pneumococci in *S. oralis* and *S. mitis* strains from primates. While NanBC is common among *S. pneumoni*ae and is missing in all *S. oralis*, all *S. oralis* contain a ß-N-acetyl-hexosaminidase which vice versa is missing in *S. pneumoniae*. The occurrence of *S. oralis* also in free-living chimpanzees suggests the assumption, that this species is part of the commensal flora of these Old-World monkeys unlike *S. pneumoniae* which has evolved with its human host. Compared to S. *pneumoniae*, *S. oralis* shows an amazing variability in factors important for biosynthesis of peptidoglycan and teichoic acid (PBP, MurMN, *lic*-cluster). Some streptococci contain a second PGP3 homologue. Additional analyses with further isolates, especially of wild animals, are necessary to determine host-specific components.

6.1 Zusammenfassung

Durch immer bessere, schnellere und auch neue Technologien steigt die Zahl der Genomsequenzierungen stetig und rapide an. Folglich besteht ein hoher Anspruch auf Automatisierung und Vereinheitlichung der nachgelagerten Verarbeitungs- und Analyseschritte der entstehenden enormen Datenmengen. In der vorliegenden Arbeit werden Genome verschiedener Streptokokken-Klone, -Stämme und -Arten miteinander verglichen, die mit verschiedenen Techniken und Methoden sequenziert, annotiert und analysiert wurden. Für die Sequenzierung wurden 454- und Illumina-Technologie verwendet. Die Assemblierung der Genome erfolgte hauptsächlich mit dem *gsAssembler* (*Newbler*) von Roche, die Annotation mit Hilfe der Annotations-Pipeline *RAST*, dem Transfertool *RATT* oder manuell. Hinsichtlich der Analysen wurden Sätze abgeleiteter Proteine verschiedener Genome miteinander verglichen und gemeinsame Komponenten, das sogenannte Core-Genom, der verwendeten Genome einer oder eng verwandter Spezies ermittelt. Für die Genome von Stämmen zweier Klone wurden detaillierte vergleichende Analysen zur Erfassung von „single nucleotide variants“ (SNV) in den Genen durchgeführt.

Fokus dieser Arbeit ist der pathogene Organismus *Streptococcus pneumoniae*. Dieser ist ein Musterbeispiel für Transformierbarkeit, aber auch für Virulenz, Pathogenität und Resistenzmechanismen gegen Antibiotika. Seine nächsten Verwandten, *S. mitis*, *S. pseudopneumoniae* und *S. oralis*, besitzen nicht so ein hohes Pathogenitätspotential wie *S. pneumoniae* und sind daher von großem Interesse, um die Evolution von *S. pneumoniae* zu verstehen. Stämme zweier *S. pneumoniae*-Klone wurden herausgegriffen. In einem Fall handelt es sich um einen Klon ST10523, der außergewöhnlich lange mit Patienten assoziiert war, die an cystischer Fibrose erkrankt waren. Diesem Klon fehlte offenbar eine aktive Hyaluronidase, einer der Hauptvirulenzfaktoren. Das Fehlen zweier Prophagencluster trug möglicherweise ebenfalls zu dem langen Verbleiben im menschlichen Wirt bei. Der Klon ST226 weist eine hohe Penizillinresistenz auf, ein Stamm ist allerdings interessanterweise sensitiv gegenüber Penicillin. Hier zeigte sich, dass die Penizillinresistenz hauptsächlich vom Vorhandensein von Mosaik-PBPs herrührte, wobei spezielle Allele von MurM und CiaH, beides Gene, die mit Penizillinresistenz in Verbindung gebracht werden, sowohl in resistenten als auch in dem sensitiven Stamm vorhanden waren. Penizillinresistenz von *S. pneumoniae* ist das Resultat von inter-spezies Gentransfer, wobei DNS nahe verwandter Streptokokken, vor allem

von *S. mitis* aber auch *S. oralis*, als Donor dient. Der Transfer von DNS vom hochgradig penizillinresistenten *S. oralis*-Stamm Uo5 auf den sensitiven *S. pneumoniae*-Stramm R6 sollte ermitteln, welche Mengen DNS dabei übertragen werden können und ob es möglich ist, das hohe Resistenzniveau des *S. oralis*-Stammes zu erreichen. Insgesamt wurde nach drei Transformationsschritten fast 19 kb *S. oralis* DNS übertragen, ungefähr 10fach weniger als im Falle von dem mit *S. pneumoniae* näher verwandten *S. mitis* als Donor. MurE wurde als neue Resistenzdeterminante identifiziert. Da das Resistenzniveau des Donorstammes in dem Rezipienten nicht erreicht werden konnte, besteht die Vermutung, dass es noch weitere, bislang unbekannte, Faktoren gibt, die zur Penizillinresistenz beitragen. Die Vergleiche von *S. pneumoniae* und seinen nahen Verwandten wurden auf Basis der abgeleiteten Proteinsequenzen durchgeführt. Den vier Genomen von *S. pneumoniae* R6, *S. pseudopneumoniae* IS7493, *S. mitis* B6 und *S. oralis* Uo5, die alle vollständig vorliegen, sind 1.041 homologe Proteine gemeinsam. Die meisten Virulenz- und Pathogenitätsfaktoren, die für *S. pneumoniae* beschrieben wurden, konnten auch in den kommensalen Spezies gefunden werden. Diese Beobachtungen wurden später durch Kilian *et al.* (Kilian, et al., 2019) durch weitere Untersuchungen bestätigt. Bei Hinzuziehen von allen 26 kompletten *S. pneumoniae* Genomen konnten nur 104 Genprodukte spezifisch für diese Spezies identifiziert werden. Untersuchungen einer größeren Anzahl verwandter Streptokokken, die aus Menschen und verschiedenen Primaten isoliert wurden, bestätigten, dass die meisten Virulenzfaktoren, die in menschlichen Pneumokokken vorhanden sind, auch in *S. mitis* und *S. oralis* aus Primaten vorkommen. Während in *S. pneumoniae* häufig NanBC vorkommt, die in allen *S. oralis* fehlt, besaßen alle *S. oralis* eine ß-N-Acetyl-Hexosaminidase, die wiederum in *S. pneumoniae* fehlt. Die Beobachtung, dass *S. oralis* auch in freilebenden Schimpansen gefunden werden konnte, legt die Vermutung nahe, dass diese Spezies Teil der kommensalen Flora dieser Altweltaffen ist und nicht, wie *S. pneumoniae*, mit dem Menschen evolviert ist. Verglichen mit *S. pneumoniae* zeigten *S. oralis* eine erstaunliche Variabilität in Faktoren, die für die Biosynthese von Peptidoglycan und Teichonsäure verantwortlich sind (PBP, MurMN und das *lic*-Cluster). Einige Streptokokken wiesen ein zweites Homolog von PBP3 auf. Weiterführende Studien mit mehr Isolaten, vor allem von freilebenden Tieren, sind notwendig, um wirtsspezifische Komponenten aufzuzeigen.

Abbreviations

aa	amino acid	NGS	next generation sequencing
ACT	Artemis Comparison Tool	N, A, C, G, T	unknown/ambiguous nucleotide, adenine, cytosine, guanine, thymine
ATP	adenosine triphosphate	NCBI	National Center for Biotechnology Information
BLAST	Basic Local Alignment and Search Tool	nt	nucleotide
Blp	Bacteriocin-like peptide	ONT	Oxford nanopore technology
CBP	choline-binding protein	PBP	penicillin binding protein
C/P/O	Cefotaxime/Piperacillin/Oxacillin	PG	peptidoglycan
CDS	(protein) coding sequence	PGAP	NCBI prokaryotic genome annotation pipeline
CF	cystic fibrosis	Ply	pneumolysin
COG	Cluster of orthologous groups	Pro	Proline
CRISPR	clustered regularly interspaced palindromic repeat	RAST	rapid annotations using subsystems technology
CSP	competence stimulating peptide	RATT	rapid annotation transfer tool
ddNTP	dideoxy nucleotide triphosphate	RBS	Ribosome binding site
DIP	Deletion/Insertion polymorphism	(t/r/nc) RNA	(transfer/ribosomal/non-coding) ribonucleic acid
DNA	desoxyribonucleic acid	RUP	repeat unit of pneumococcus
EMBL	European Molecular Biology Laboratory	SFF	Standard Flowgram Format
Gln	Glutamine	SLO/SLS	Streptolysin O/S
HPN	homopolynucleotide	SLV	single locus variant
ICE	integrative and conjugative element	SNP	single nucleotide polymorphism
Indel	single nucleotide insertion and/or deletion	SNV	single nucleotide variation (SNP or indel/DIP)
IS	insertion sequence	TCS	two-component system
LTA	Lipoteichoic acid	TGS	Third generation sequencing technology
MIC	minimal inhibitory concentration	VF	virulence factor
MLSA	multi locus sequence analysis	WASABI	Web-Accessible Sequence Analysis for Biological Inference
MLST	multi locus sequence typing	WTA	wall teichoic acid
(I/H) MM	(interpolated/hidden) Markov Model		

Table index

Figure index

References

Abranches, J, et al. 2018. Biology of oral Streptococci. *Microbiol Spectr.* Oct 2018, Vol. 6, 5.

Ahn, S J, et al. 2014. Discovery of novel peptides regulating competence development in Streptococcus mutans. *J Bacteriol.* Nov 2014, Vol. 196, 21, pp. 3735-3745.

Ajay, S S, et al. 2011. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* Sep 2011, Vol. 21, 9, pp. 1498-1505.

AlonsoDeVelasco, E, et al. 1995. Streptococcus pneumoniae: virulence factors, pathogenesis, and vaccines. *Microbiol Rev.* 1995, Vol. 59, 4, pp. 591-603.

Altschul, S F, et al. 1990. Basic local alignment search tool. *Journal of molecular biology.* 1990, Vol. 215, 3, pp. 403–410.

Andam, C P and Hanage, W P. 2015. Mechanisms of genome evolution of Streptococcus. *Infect Genet Evol.* Jul 2015, Vol. 33, pp. 334-342.

Avery, O T and Dubos, R. 1931. THE PROTECTIVE ACTION OF A SPECIFIC ENZYME AGAINST TYPE III PNEUMOCOCCUS INFECTION IN MICE. *J Exp Med.* 30 Jun 1931, Vol. 54, 1, pp. 73-89.

Avery, O T, MacLeod, C M and McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med.* 1 Feb 1944, Vol. 79, 2, pp. 137-158.

Aziz, R K, et al. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics.* 9, 8 Feb 2008, p. 75ff.

Babina, A M, et al. 2015. An S6:S18 complex inhibits translation of E. coli rpsF. *RNA.* Dec 2015, Vol. 21, 12, pp. 2039-2046.

Bakkali, M. 2013. Could DNA uptake be a side effect of bacterial adhesion and twitching motility? *Arch Microbiol.* Apr 2013, Vol. 195, 4, pp. 279–289.

Balachandran, P, et al. 2001. The autolytic enzyme LytA of Streptococcus pneumoniae is not responsible for releasing pneumolysin. *J Bacteriol.* May 2001, Vol. 183, 10, pp. 3108-3116.

Barnard, J P and Stinson, M W. 1996. The alpha-hemolysin of Streptococcus gordonii is hydrogen peroxide. *Infect Immun.* Sep 1996, Vol. 64, 9, pp. 3853-7.

Baron, E J. 1996. Classification. [book auth.] S Baron. *Baron's Medical Microbiology.* 4th edition. Galveston (TX) : University of Texas Medical Branch at Galveston, 1996.

Bean, B and Tomasz, A. 1977. Choline metabolism in pneumococci. *J Bacteriol.* Apr 1977, Vol. 130, 1, pp. 571-574.

Bentley, D R, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008, Vol. 456, 7218, pp. 53–59.

Bentley, S D, et al. 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet.* Mar 2006, Vol. 2, 3, p. e31.

Benton, K A, Paton, J C and Briles, D E. 1997. Differences in virulence for mice among Streptococcus pneumoniae strains of capsular types 2, 3, 4, 5, and 6 are not attributable to differences in pneumolysin production. *Infect Immun.* Apr 1997, Vol. 65, 4, pp. 1237-1244.

Bhakdi, S, Tranum-Jensen, J and Sziegoleit, A. 1985. Mechanism of membrane damage by streptolysin-O. *Infect Immun.* Jan 1985, Vol. 47, 1, pp. 52-60.

Bishop, C J, et al. 2009. Assigning strains to bacterial species via the internet. *BMC Biol.* 26 Jan 2009, Vol. 7, p. 3.

Blake, F G. 1916. The formation of methemoglobin by Streptococcus viridans. *J Exp Med.* 1 Oct 1916, Vol. 24, 4, pp. 315-327.

Boldogkői, Z, et al. 2019. Long-Read Sequencing - A Powerful Tool in Viral Transcriptome Research. *Trends Microbiol.* Jul 2019, Vol. 27, 7, pp. 578-592.

Bracco, R M, et al. 1957. Transformation reactions between Pneumococcus and three strains of Streptococci. *J Exp Med.* 1 Aug 1957, Vol. 106, 2, pp. 247–259.

Brooks-Walter, A, Briles, D E and Hollingshead, S K. 1999. The pspC gene of Streptococcus pneumoniae encodes a polymorphic protein, PspC, which elicits cross-reactive antibodies to PspA and provides immunity to pneumococcal bacteremia. *Infect Immun.* Dec 1999, Vol. 67, 12, pp. 6533-6542.

Brown, J S, et al. 2002. Characterization of pit, a Streptococcus pneumoniae iron uptake ABC transporter. *Infect Immun.* Aug 2002, Vol. 70, 8, pp. 4389-4398.

Brown, J S, et al. 2001. Immunization with components of two iron uptake ABC transporters protects mice against systemic Streptococcus pneumoniae infection. *Infect Immun.* Nov 2001, Vol. 69, 11, pp. 6702-6706.

Brown, J S, Gilliland, S M and Holden, D W. 2001. A Streptococcus pneumoniae pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol Microbiol.* May 2001, Vol. 40, 3, pp. 572-585.

Brückner, R, et al. 2004. Mosaic genes and mosaic chromosomes-genomic variation in Streptococcus pneumoniae. *Int J Med Microbiol.* Sep 2004, Vol. 294, 2-3, pp. 157-168.

Brueggemann, A B, et al. 2017. Pneumococcal prophages are diverse, but not without structure or history. *Sci Rep.* 20 Feb 2017, Vol. 7, p. 42976.

Bubunenko, M, Baker, T and Court, D L. 2007. Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in Escherichia coli. *J Bacteriol.* Apr 2007, Vol. 189, 7, pp. 2844-2853.

Burnside, K, et al. 2010. Regulation of hemolysin expression and virulence of Staphylococcus aureus by a serine/threonine kinase and phosphatase. *PLoS One.* Jun 2010, Vol. 5, 6, p. e11071.

Byrd, V S and Nemeth, A S. 2017. A case of infective endocarditis and spinal epidural abscess caused by Streptococcus mitis bacteremia. *Case Rep Infect Dis.* 2017, Vol. 2017, p. 7289032.

Calidas, D, Lyon, H and Culver, G M. 2014. The N-terminal extension of S12 influences small ribosomal subunit assembly in Escherichia coli. *RNA.* Mar 2014, Vol. 20, 3, pp. 321-330.

Carr A, Sledjeski DD, Podbielski A, Boyle MD, Kreikemeyer B. 2001. Similarities between complement-mediated and streptolysin S-mediated hemolysis. *J Biol Chem.* 9 Nov 2001, Vol. 276, 45, pp. 41790-6.

Carver, T J, et al. 2005. ACT, The Artemis Comparison Tool. *Bioinformatics (Oxford, England).* 2005, Vol. 21, 16, pp. 3422–3423.

Carver, T, et al. 2012. Artemis. An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England).* 2012, Vol. 28, 4, pp. 464–469.

Casadevall, A and Pirofski, L A. 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infect Immun.* Aug 1999, Vol. 67, 8, pp. 3703-3713.

Chaisson, M J and Pevzner, P A. 2008. Short read fragment assembly of bacterial genomes. *Genome Res.* Feb 2008, Vol. 18, 2, pp. 324-330.

Chambers, H F. 1999. Penicillin-binding protein-mediated resistance in pneumococci and staphylococci. *J Infect Dis.* Mar 1999, Vol. 179 Suppl 2, pp. S353-9.

Chi, Y C, et al. 2017. Streptococcus pneumoniae IgA1 protease: A metalloprotease that can catalyze in a split manner in vitro. *Protein Sci.* Mar 2017, Vol. 26, 3, pp. 600-610.

Claverys, J P, Martin, B and Polard, P. 2009. The genetic transformation machinery: composition, localization, and mechanism. *FEMS Microbiol Rev.* May 2009.

Cock, P J, et al. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* Apr 2010, Vol. 38, 6, pp. 1767-1771.

Coffey, T J, et al. 1991. Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of Streptococcus pneumoniae. *Mol Microbiol.* Sep 1991, Vol. 5, 9, pp. 2255-2260.

Coffey, T J, et al. 1998. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of Streptococcus pneumoniae. *Mol Microbiol.* Jan 1998, Vol. 27, 1, pp. 73-83.

Coffey, T J, et al. 1998. Serotype 19A variants of the Spanish serotype 23F multiresistant clone of Streptococcus pneumoniae. *Microb Drug Resist.* 1998, Vol. 4, 1, pp. 51-55.

Cole, J N, et al. 2008. Human pathogenic streptococcal proteomics and vaccine. *Proteomics Clin Appl.* 2008, Vol. 2, 3, pp. 387–410.

Collins, F S and Weissman, S M. 1984. Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. *Proc Natl Acad Sci U S A.* Nov 1984, Vol. 81, 21, pp. 6812-6816.

Cowley, L A, et al. 2018. Evolution via recombination: Cell-to-cell contact facilitates larger recombination events in Streptococcus pneumoniae. *PLoS Genet.* 13 Jun 2018, Vol. 14, 6, p. e1007410.

Croucher, N J, et al. 2011. Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genomics.* 18 Feb 2011, Vol. 12, p. 120.

Croucher, N J, et al. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet.* Jun 2013, Vol. 45, 6, pp. 656-663.

Croucher, N J, et al. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science.* 28 Jan 2011, Vol. 331, 6016, pp. 430-434.

Croucher, N J, et al. 2009. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniae Spain 23F ST81. *J Bacteriol.* Mar 2009, Vol. 191, 5, pp. 1480-1489.

Czajkowsky, D M, et al. 2004. Vertical collapse of a cytolysin prepore moves its transmembrane beta-hairpins to the membrane. *EMBO J.* 18 Aug 2004, Vol. 23, 16, pp. 3206-3215.

Deibel, R H and Seeley, H W Jr. 1974. Family II: Streptococcuceae. Fam. nov. [book auth.] D H Bergey, R E Buchanan and N E Gibbons. *Bergy's manual of determinative bacteriology.* 8th ed. s.l. : Baltimore: Williams & Wilkins, 1974, pp. 450-517.

Delcher, A L, et al. 1999. Alignment of whole genomes. *Nucleic Acids Res.* Jun 1999, Vol. 27, 11, pp. 2369-2376.

Delcher, A L, et al. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 1 Jun 2002, Vol. 30, 11, pp. 2478-2483.

Denapaite, D, et al. 2010. The genome of Streptococcus mitis B6-what is a commensal? *PLoS ONE.* 2010, Vol. 5, 2, p. e9426.

Denpaite, D, et al. 2016. Highly variable Streptococcus oralis strains are common among viridans Streptococci Isolated from primates. *mSphere.* 9 Mar 2016, Vol. 1, 2, pp. e00041-15.

Dintilhac, A, et al. 1997. Competence and virulence of Streptococcus pneumoniae: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. *Mol Microbiol.* Aug 1997, Vol. 25, 4, pp. 727-739.

Dochez, A R and Avery, O T. 1917. The elaboration of specific soluble substance by pneumococcus during growth. *J Exp Med.* 1 Oct 1917, Vol. 26, 4, pp. 477-493.

Dohm, J C, et al. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Sep 2008, Vol. 36, 16, p. e105.

Donati, C, et al. 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol.* 2010, Vol. 11, 10, p. R107.

Dowson, C G, Hutchinson, A and Spratt, B G. 1989. Extensive re-modelling of the transpeptidase domain of penicillin-binding protein 2B of a penicillin-resistant South African isolate of Streptococcus pneumoniae. *Mol Microbiol.* Jan 1989, Vol. 3, 1, pp. 95-102.

Drijkoningen, J J and Rohde, G G. 2014. Pneumococcal infection in adults: burden of disease. *Clin Microbiol Infect.* May 2014, Vol. 20 Suppl 5, pp. 45-51.

Eisen, J A, et al. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 2000, Vol. 1, 6, p. RESEARCH0011.

Engholm, D H, et al. 2017. A visual review of the human pathogen Streptococcus pneumoniae. *FEMS Microbiol Rev.* 1 Nov 2017, Vol. 41, 6, pp. 854-879.

Enright, M C and Spratt, B G. 1999. Multilocus sequence typing. *Trends Microbiol.* Dec 1999, Vol. 7, 12, pp. 482-487.

Facklam, R. 2002. What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. *Clin Microbiol Rev.* Oct 2002, Vol. 15, 4, pp. 613-630.

Fang, H, et al. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.* 2014, Vol. 6, 10, p. 89.

Fani, F, et al. 2014. Genomic analyses of DNA transformation and penicillin resistance in Streptococcus pneumoniae clinical isolates. *Antimicrob Agents Chemother.* 2014, Vol. 58, 3, pp. 1397-1403.

Felsenstein, J. 2013. PHYLIP Phylogeny Inference Package Version 3.695. [Online] Apr 2013. http://evolution.genetics.washington.edu/phylip/doc/main.html.

Fernandes, G R, et al. 2017. Genomic comparison among lethal invasive strains of Streptococcus pyogenes serotype M1. *Front Microbiol.* 23 Oct 2017, Vol. 8, p. 1993.

Filipe, S R and Tomasz, A. 2000. Inhibition of the expression of penicillin resistance in Streptococcus pneumoniae by inactivation of cell wall muropeptide branching genes. *Proc Natl Acad Sci U S A.* 25 Apr 2000, Vol. 97, 9, pp. 4891-4896.

Fischer, W. 1997. Pneumococcal lipoteichoic and teichoic acid. *Microb Drug Resist.* 1997, Vol. 3, 4, pp. 309-325.

Fleischmann, R D, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science.* 28 Jul 1995, Vol. 296, 5223, pp. 496-512.

Fraenkel, A. 1886b. Weitere Beitrage zur Lehre von den Mikrococcen der genuinen fibrinosen Pneumonie. *Zeitschrift für Klinische Medicin.* 1886b, 11, pp. 437-458.

Fullwood, Melissa J., et al. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research.* Apr 2009, Vol. 19, 4, pp. 521–532.

Gardete, S, et al. 2004. Role of murE in the Expression of beta-lactam antibiotic resistance in Staphylococcus aureus. *J Bacteriol.* Mar 2004, Vol. 186, 6, pp. 1705-1713.

Garriss, G, et al. 2019. Genomic characterization of the emerging pathogen Streptococcus pseudopneumoniae. *MBio.* 25 Jun 2019, Vol. 10, 3, pp. e01286-19.

Geno, K A, et al. 2015. Pneumococcal Capsules and Their Types: Past, Present, and Future. *Clin Microbiol Rev.* Jul 2015, Vol. 28, 3, pp. 871-899.

Gertz, E M, et al. 2006. Composition-based statistics and translated nucleotide searches. *BMC Biology.* 2006, Vol. 4, p. 41.

Hakenbeck R, König A, Kern I, van der Linden M, Keck W, Billot-Klein D, Legrand R, Schoot B, Gutmann L. 1998. Acquisition of five high-Mr penicillin-binding protein variants during transfer of high-level beta-lactam resistance from Streptococcus mitis to Streptococcus pneumoniae. *J Bacteriol.* Apr 1998, Vol. 180, 7, pp. 1831-1840.

Hakenbeck, R, et al. 1999. beta-lactam resistance in Streptococcus pneumoniae: penicillin-binding proteins and non-penicillin-binding proteins. *Mol Microbiol.* Aug 1999, Vol. 33, 4, pp. 673-678.

Hakenbeck, R, et al. 2012. Molecular mechanisms of β-lactam resistance in Streptococcus pneumoniae. *Future Microbiol.* Mar 2012, Vol. 7, 3, pp. 395-410.

Hakenbeck, R, et al. 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of Streptococcus pneumoniae. *Infect Immun.* Apr 2001, Vol. 69, 4, pp. 2477-2486.

Halfmann A, Schnorpfeil A, Müller M, Marx P, Günzler U, Hakenbeck R, Brückner R. 2011. Activity of the two-component regulatory system CiaRH in Streptococcus pneumoniae R6. *J Mol Microbiol Biotechnol.* Apr 2011, Vol. 20, 2, pp. 96-104.

Hammerschmidt, S, et al. 1999. Identification of pneumococcal surface protein A as a lactoferrin-binding protein of Streptococcus pneumoniae. *Infect Immun.* Apr 1999, Vol. 67, 4, pp. 1683-1687.

Hammerschmidt, S, et al. 1997. SpsA, a novel pneumococcal surface protein with specific binding to secretory immunoglobulin A and secretory component. *Mol Microbiol.* Sep 1997, Vol. 25, 6, pp. 1113-1124.

Han, X, et al. 2016. Improvement of the Texture of Yogurt by Use of Exopolysaccharide Producing Lactic Acid Bacteria. *Biomed Res Int.* 2016, 2016, p. 7945675.

Hava, D L and Camilli, A. 2002. Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors. *Mol Microbiol.* Sep 2002, Vol. 45, 5, pp. 1389-1406.

Heidelberger, M and Avery, O T. 1923. The soluble specific substance of pneumococcus. *J Exp Med.* 30 Jun 1923, Vol. 38, 1, pp. 73-79.

Hiller, N L and Sá-Leão, R. 2018. Puzzling over the pneumococcal pangenome. *Front Microbiol.* 30 Oct 2018, Vol. 9, p. 2580.

Hiller, N L, et al. 2007. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. *J Bacteriol.* Nov 2007, Vol. 189, 22, pp. 8186-8195.

Hillier, L W, et al. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nature methods.* 2008, Vol. 5, 2, pp. 183–188.

Hollingshead, S K, et al. 2006. Pneumococcal surface protein A (PspA) family distribution among clinical isolates from adults over 50 years of age collected in seven countries. *J Med Microbiol.* Feb 2006, Vol. 55(Pt 2), pp. 215-221.

Hong, G F. 1981. A method for sequencing single-stranded cloned DNA in both directions. *Biosci Rep.* Mar 1981, Vol. 1, 3, pp. 243-252.

Hoskins, J, et al. 2001. Genome of the bacterium Streptococcus pneumoniae strain R6. *J.Bacteriol.* 2001, Vol. 183, 19, pp. 5709–5717.

Huse, S M, et al. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 2007, Vol. 8, 7, p. R143.

Iannelli, F, Pearce, B J and Pozzi, G. 1999. The type 2 capsule locus of Streptococcus pneumoniae. *J Bacteriol.* Apr 1999, Vol. 181, 8, pp. 2652-2654.

Iannelli, Fi, Oggioni, M R and Pozzi, G. 2002. Allelic variation in the highly polymorphic locus pspC of Streptococcus pneumoniae. *Gene.* 6 Feb 2002, Vol. 284, 1-2, pp. 63-71.

Ibrahim, Y M, et al. 2004. Role of HtrA in the virulence and competence of Streptococcus pneumoniae. *Infect Immun.* Jun 2004, Vol. 72, 6, pp. 3584-3591.

Java. https://www.java.com. [Online]

Jedrzejas, M J. 2001. Pneumococcal virulence factors. *Microbiology and molecular biology reviews : MMBR.* 2001, Vol. 65, 2, pp. 187-207 ; first page, table of contents.

Johnson, C M and Grossman, A D. 2015. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu Rev Genet.* 2015, Vol. 49, pp. 577-601.

Johnston, C, et al. 2010. Detection of large numbers of pneumococcal virulence genes in streptococci of the mitis group. *J Clin Microbiol.* Aug 2010, Vol. 48, 8, pp. 2762-2769.

Johnston, C, et al. 2013. Natural genetic transformation generates a population of merodiploids in Streptococcus pneumoniae. *PLoS genetics.* 2013, Vol. 9, 9, p. e1003819.

Jonsson, S, et al. 1985. Phagocytosis and killing of common bacterial pathogens of the lung by human alveolar macrophages. *J Infect Dis.* Jul 1985, Vol. 152, 1, pp. 4-13.

Kanclerski, K and Möllby, R. 1987. Production and purification of Streptococcus pneumoniae hemolysin (pneumolysin). *J Clin Microbiol.* Feb 1987, Vol. 25, 2, pp. 222-225.

Kauff, F, Cox, C J and Lutzoni, F. 2007. WASABI: An automated sequence processing system for multigene phylogenies. *Syst. Biol.* 2007, Vol. 56, 3, pp. 523–531.

Kawamura, Y, et al. 1995. Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic relationships among members of the genus Streptococcus. *Int J Syst Bacteriol.* Apr 1995, Vol. 45, 2, pp. 406-408.

Keller, L E, Robinson, D A and McDaniel, L S. 2016. Nonencapsulated Streptococcus pneumoniae: Emergence and Pathogenesis. *mBio.* 22 Mar 2016, Vol. 7, 2, p. e01792.

Keogh, B P. 1970. Survival and activity of frozen starter cultures for cheese manufacture. *Appl Microbiol.* 19, June 1970, 6, pp. 928-931. PMC376826 Journal Article.

Khan, M N, et al. 2012. PcpA of Streptococcus pneumoniae mediates adherence to nasopharyngeal and lung epithelial cells and elicits functional antibodies in humans. *Microbes Infect.* Oct 2012, Vol. 14, 12, pp. 1102-1110.

Kilian, M and Tettelin, H. 2019. Identification of virulence-associated properties by comparative genome analysis of Streptococcus pneumoniae, S. pseudopneumoniae, S. mitis, three S. oralis subspecies, and S. infantis. *MBio.* 3 Sep 2019, Vol. 10, 5, pp. e01985-19.

Kilian, M, et al. 2008. Evolution of Streptococcus pneumoniae and its close commensal relatives. *PLoS One.* 16 Jul 2008, Vol. 3, 7, p. e2683.

Kilian, M, et al. 2014. Parallel evolution of Streptococcus pneumoniae and Streptococcus mitis to pathogenic and mutualistic lifestyles. *mBio.* 2014, Vol. 5, 4, pp. e01490-14.

Kisand, V and Lettieri, T. 2013. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC Genomics.* Apr 2013, Vol. 14, p. 211.

Klein, D L. 1999. Pneumococcal disease and the role of conjugate vaccines. *Microbial drug resistance (Larchmont, N.Y.).* 1999, Vol. 5, 2, pp. 147–157.

Klugman, K P. 2002. The successful clone: the vector of dissemination of resistance in Streptococcus pneumoniae. *J Antimicrob Chemother.* Dec 2002, Vol. 50, pp. Suppl S2:1-5.

Krzyściak W, Pluskwa KK, Jurczak A, Kościelniak D. 2013. The pathogenicity of the Streptococcus genus. *Eur J Clin Microbiol Infect Dis.* Nov 2013, Vol. 32, 11, pp. 1361-1376.

Kurtz, S, et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 2004, Vol. 5, 2, p. R12.

Laible, G, et al. 1989. Nucleotide sequences of the pbpX genes encoding the penicillin-binding proteins 2x from Streptococcus pneumoniae R6 and a cefotaxime-resistant mutant, C506. Oct 1989, Vol. 3, 10, pp. 1337-1347.

Laible, G, Spratt, B G and Hakenbeck, R. 1991. Interspecies recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of Streptococcus pneumoniae. *Mol Microbiol.* Aug 1991, Vol. 5, 8, pp. 1993-2002.

Lancefield, R C. 1933. A serological differentiation of human and other groups of hemolytic Streptococci. *The Journal of Experimental Medicine.* 1933, Vol. 57, 4, pp. 571-595.

Land, M, et al. 2015. Insights from 20 years of bacterial genome sequencing. Mar 2015, Vol. 15, 2, pp. 141-161.

Lanie, J A, et al. 2007. Genome sequence of Avery's virulent serotype 2 strain D39 of Streptococcus pneumoniae and comparison with that of unencapsulated laboratory strain R6. *J Bacteriol.* Jan 2007, Vol. 189, 1, pp. 38-51.

Lau, G W, et al. 2001. A functional genomic analysis of type 3 Streptococcus pneumoniae virulence. *Mol Microbiol.* May 2001, Vol. 40, 3, pp. 555-571.

Laux A, Sexauer A, Sivaselvarajah D, Kaysen A, Brückner R. 2015. Control of competence by related non-coding csRNAs in Streptococcus pneumoniae R6. *Front Genet.* 20 Jul 2015, Vol. 20, 2, pp. 96-104.

Levin, B R and Cornejo, O E. 2009. The population and evolutionary dynamics of homologous gene recombination in bacterial populations. *PLoS Genet.* Aug 2009, Vol. 5, 8, p. e1000601.

Li, H, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 15 Aug 2009, Vol. 25, 16, pp. 2078-2079.

Li, H, Ruan, J and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* Nov 2008, Vol. 18, 11, pp. 1851-1858.

Li, J, et al. 2007. PspA and PspC minimize immune adherence and transfer of pneumococci from erythrocytes to macrophages through their effects on complement activation. *Infect Immun.* Dec 2007, Vol. 75, 12, pp. 5877-5885.

Liu, L, et al. 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology.* 2012, Vol. 2012, p. 251364.

Liu, L, et al. 2011. Microbial production of hyaluronic acid: current state, challenges, and perspectives. *Microb Cell Fact.* 16 Nov 2011, Vol. 10, p. 99.

Loughran, A J, Orihuela, C J and Tuomanen, E I. 2019. Streptococcus pneumoniae: Invasion and inflammation. *Microbiol Spectr.* Mar 2019, Vol. 7, 2.

Lu, H, Giordano, F and Ning, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, proteomics & bioinformatics.* 2016, Vol. 14, 5, pp. 265–279.

Luo, C, et al. 2012. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE.* 2012, Vol. 7, 2.

MacLeod, C M and Roe, A S. 1956. Effect of silicate on Gram staining and viability of pneumococci and other bacteria. *The Journal of experimental medicine.* 1956, Vol. 103, 4, pp. 453–463.

Maddison, D R, Swofford, D L and Maddison, W P. 1997. NEXUS. *Syst.Biol.* 1997, Vol. 46, 4, pp. 590–621.

Madhour, A., Maurer, P. and Hakenbeck, R. 2011. Cell surface proteins in S. pneumoniae, S. mitis and S. oralis. *Iranian journal of microbiology.* 2011, Vol. 3, 2, pp. 58–67.

Madigan, M T, Martinko, J M and Parker, J. 2002. Bacterial genetics. *Brock - Biology of microorganisms.* Tenth Edition. s.l. : Prentice Hall International, 2002, pp. 264-320.

—. 2002. Human-microbe interactions. [book auth.] Michael T Madigan, John M Martinko and Jack Parker. *Brock - Biology of microorganisms.* Tenth Edition. s.l. : Prentice Hall International, 2002, pp. 727-754.

Mahillon, J and Chandler, M. 1998. Insertion sequences. *Microbiology and molecular biology reviews : MMBR.* 1998, Vol. 62, 3, pp. 725–774.

Manso, A S, et al. 2014. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nature communications.* 2014, Vol. 5, p. 5055.

Marçais, G, et al. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol.* 26 Jan 2018, Vol. 14, 1, p. e1005944.

Margulies, M, et al. 2005. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature.* 2005, Vol. 437, 7057, pp. 376–380.

Marmorek, A. 1895. *Ann. Inst. Pasteur.* 1895, Vol. 9, 523.

Marx, P, et al. 2010. Identification of genes for small non-coding RNAs that belong to the regulon of the two-component regulatory system CiaRH in Streptococcus. *BMC Genomics.* 24 Nov 2010, Vol. 11, p. 661.

McKenna, A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* Sep 2010, Vol. 20, 9, pp. 1297-1303.

Meiers, M. 2015. Genetische Analyse von Resistenzdeterminanten in Streptococcus pneumoniae. *Dissertation.* 2015.

Metzker, M L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet.* Jan 2010, Vol. 11, 1, pp. 31-46.

Miller, J R, Koren, S and Sutton, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics.* Jun 2010, Vol. 95, 6, pp. 315-327.

Mitchell, A M and Mitchell, T J. 2010. Streptococcus pneumoniae: virulence factors and variation. *Clin Microbiol Infect.* May 2010, Vol. 16, 5, pp. 411-418.

Molloy, E M, et al. 2015. Identification of the minimal cytolytic unit for streptolysin S and an expansion of the toxin family. *BMC Microbiol.* 24 Jul 2015, 15.

Müller, M, et al. 2011. Effect of new alleles of the histidine kinase gene ciaH on the activity of the response regulator CiaR in Streptococcus pneumoniae R6. *Microbiology (Reading, England).* 2011, Vol. 157, Pt 11, pp. 3104–3112.

Munita, J M and Arias, C A. 2016. Mechanisms of antibiotic resistance. *Microbiol Spectr.* Apr 2016, Vol. 4, 2.

Muñoz-López, M and García-Pérez, J L. 2010. DNA Transposons: Nature and Applications in Genomics. *Curr Genomics.* Apr 2010, Vol. 11, 2, pp. 115-128.

Nakagawa, I, et al. 2003. Genome sequence of an M3 strain of Streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res.* Jun 2003, Vol. 13, 6A, pp. 1042-1055.

NCBI. Gapped format for genome submissions. *Gapped format for genome submissions.* [Online] National Center for Biotechnology Information, U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/genbank/wgs_gapped.

Needleman, S B and Wunsch, C D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology.* 1970, Vol. 48, 3, pp. 443–453.

Neeleman, C, et al. 2004. Pneumolysin is a key factor in misidentification of macrolide-resistant Streptococcus pneumoniae and is a putative virulence factor of S. mitis and other streptococci. *J Clin Microbiol.* Sep 2004, Vol. 42, 9, pp. 4355-4357.

Oggioni, M R and Claverys, J P. 1999. Repeated extragenic sequences in prokaryotic genomes. *Microbiology (Reading, England).* 1999, Vol. 145 (Pt 10), pp. 2647–2653.

Ogunniyi, A D, et al. 2010. Central role of manganese in regulation of stress responses, physiology, and metabolism in Streptococcus pneumoniae. *J Bacteriol.* Sep 2010, Vol. 192, 17, pp. 4489-4497.

Ogunniyi, A D, Giammarinaro, P and Paton, J C. 2002. The genes encoding virulence-associated proteins and the capsule of Streptococcus pneumoniae are upregulated and differentially expressed in vivo. *Microbiology.* Jul 2002, Vol. 148(Pt 7), pp. 2045-2053.

Orihuela, C J, et al. 2004. Tissue-specific contributions of pneumococcal virulence factors to pathogenesis. *J Infect Dis.* 1 Nov 2004, Vol. 190, 9, pp. 1661-1669.

Otto, T D, et al. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* May 2011, Vol. 39, 9, p. e57.

Park, I H, et al. 2007. Discovery of a new capsular serotype (6C) within serogroup 6 of Streptococcus pneumoniae. *J Clin Microbiol.* Apr 2007, Vol. 45, 4, pp. 1225-1233.

Pasta, F and Sicard, M A. 1999. Polarity of recombination in transformation of Streptococcus pneumoniae. *Proc Natl Acad Sci U S A.* 16 Mar 1999, Vol. 96, 6, pp. 2943-8.

Pasteur, L. 1881. Sur une maladie nouvelle provoquée par la salive d'un enfant mort de rage. *Comptes rendus de l'Académie des Sciences de Paris.* 1881, Vol. 92, 159.

Paton, J C, Berry, A M and Lock, R A. 1997. Molecular analysis of putative pneumococcal virulence proteins. *Microb Drug Resist.* 1997, Vol. 3, 1, pp. 1-10.

Philippe, H and Douady, C J. 2003. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol.* Oct 2003, Vol. 6, 5, pp. 498-505.

Polissi, A, et al. 1998. Large-scale identification of virulence genes from Streptococcus pneumoniae. *Infect Immun.* Dec 1998, Vol. 66, 12, pp. 5620-5629.

Price, K E and Camilli, A. 2009. Pneumolysin localizes to the cell wall of Streptococcus pneumoniae. *J Bacteriol.* Apr 2009, Vol. 191, 7, pp. 2163-2168.

Prlic, A, et al. 2012. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics.* 2012, Vol. 28, 20, pp. 2693–2695.

Ravin, A W. 1960. THE ORIGIN OF BACTERIAL SPECIES: Genetic Recombination and Factors Limiting It Between Bacterial Populations. *Bacteriol Rev.* Jun 1960, Vol. 24, 2, pp. 201–220.

Reichmann, P, et al. 1997. A global gene pool for high-level cephalosporin resistance in commensal Streptococcus species and Streptococcus pneumoniae. *J Infect Dis.* Oct 1997, Vol. 176, 4, pp. 1001-1012.

Reichmann, P, et al. 2011. Genome of Streptococcus oralis strain Uo5. *J.Bacteriol.* 2011, Vol. 193, 11, pp. 2888–2889.

Reichmann, P, et al. 1995. Penicillin-resistant Streptococcus pneumoniae in Germany: : genetic relationship to clones from other European countries. *J.Med.Microbiol.* 1995, Vol. 43, 5, pp. 377–385.

Reinert, R R. 2009. The antimicrobial resistance profile of Streptococcus pneumoniae. *Clin Microbiol Infect.* Apr 2009, Vol. 15 Suppl 3, pp. 7-11.

Rieger, M, et al. 2017. Draft genome sequences of two Streptococcus pneumoniae serotype 19A sequence Type 226 clinical isolates from Hungary, Hu17 with High-Level Beta-Lactam Resistance and Hu15 of a penicillin-sensitive phenotype. *Genome Announc.* 18 May 2017, Vol. 5, 20, pp. 000401-17.

Rieger, M, Mauch, H and Hakenbeck, R. 2017. Long persistence of a Streptococcus pneumoniae 23F clone in a cystic fibrosis patient. *mSphere.* 7 Jun 2017, Vol. 2, 3, pp. e00201-17.

Ring, A, Weiser, J N and Tuomanen, E I. 1998. Pneumococcal trafficking across the blood-brain barrier. Molecular analysis of a novel bidirectional pathway. *J Clin Invest.* 15 Jul 1998, Vol. 102, 2, pp. 347-360.

Rissman, A I, et al. 2009. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics.* 15 Aug 2009, Vol. 25, 16, pp. 2071-2073.

Roberts, A P and Mullany, P. 2011. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol Rev.* Sep 2011, Vol. 35, 5, pp. 856-871.

Rosenow, C, et al. 1997. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of Streptococcus pneumoniae. *Mol Microbiol.* Sep 1997, Vol. 25, 5, pp. 819-829.

Roy, D, et al. 2014. Role of the capsular polysaccharide as a virulence factor for Streptococcus suis serotype 14. *Can J Vet Res.* Apr 2014, Vol. 79, 2, pp. 141-146.

Salles, C, et al. 1992. The high level streptomycin resistance gene from Streptococcus pneumoniae is a homologue of the ribosomal protein S12 gene from Escherichia coli. *Nucleic Acids Res.* 25 Nov 1992, Vol. 20, 22, p. 6103.

Salvadori, G, et al. 2019. Competence in Streptococcus pneumoniae and close commensal relatives: Mechanisms and implications. *Front Cell Infect Microbiol.* 3 Apr 2019, Vol. 9, p. 94.

Salzberg, S L, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* Mar 2012, Vol. 22, 3, pp. 557-567.

Sampson, J S, et al. 1997. Limited diversity of Streptococcus pneumoniae psaA among pneumococcal vaccine serotypes. *Infect Immun.* May 1997, Vol. 65, 5, pp. 1967-1971.

Sauerbier, J, et al. 2012. Streptococcus pneumoniae R6 interspecies transformation: genetic analysis of penicillin resistance determinants and genome-wide recombination events. *Mol Microbiol.* Nov 2012, Vol. 86, 3, pp. 692-706.

Scheffers, D J and Pinho, M G. 2005. Bacterial cell wall synthesis: new insights from localization studies. *Microbiol Mol Biol Rev.* Dec 2005, Vol. 69, 4, pp. 585-607.

Schnorpfeil, A, et al. 2013. Target evaluation of the non-coding csRNAs reveals a link of the two-component regulatory system CiaRH to competence control in Streptococcus pneumoniae R6. *Molecular microbiology.* 2013, Vol. 89, 2, pp. 334–349.

Schroeder, M R and Stephens, D S. 2016. Macrolide resistance in Streptococcus pneumoniae. *Front Cell Infect Microbiol.* 21 Sep 2016, Vol. 6, p. 98.

Schweizer, I, et al. 2017. New Aspects of the Interplay between Penicillin Binding Proteins, murM, and the Two-Component System CiaRH of Penicillin-Resistant Streptococcus pneumoniae Serotype 19A Isolates from Hungary. *Antimicrobial Agents and Chemotherapy.* 2017, Vol. 61, 7.

Sebert, M E, et al. 2002. Microarray-based identification of htrA, a Streptococcus pneumoniae gene that is regulated by the CiaRH two-component system and contributes to nasopharyngeal colonization. *Infect Immun.* Aug 2002, Vol. 70, 8, pp. 4059-4067.

Shahinas, D, et al. 2013. Comparative genomic analyses of Streptococcus pseudopneumoniae provide insight into virulence and commensalism dynamics. *PLoS One.* 19 Jun 2013, Vol. 8, 6, p. e65670.

Shahinas, D., et al. 2011. Whole-genome sequence of Streptococcus pseudopneumoniae isolate IS7493. *J.Bacteriol.* 2011, Vol. 193, 21, pp. 6102–6103.

Shaper, M, et al. 2004. PspA protects Streptococcus pneumoniae from killing by apolactoferrin, and antibody to PspA enhances killing of pneumococci by apolactoferrin [corrected]. *Infect Immun.* Sep 2004, Vol. 72, 9, pp. 5031-5040.

Shottmuller, H. Die Artunterscheidung der fur den menschen Pathogen Streptokokken durch Blutagar. *Munch. Med. Wochenschr.* 50, pp. 849-853.

Sievers, F and Higgins, D G. 2014. Clustal omega. *Curr Protoc Bioinformatics.* 12 12 2014, Vol. 43, pp. 3.13.1-16.

Sims, D, et al. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet.* Feb 2014, Vol. 15, 2, pp. 121-132.

Sjöström, K, et al. 2006. Clonal and capsular types decide whether pneumococci will act as a primary or opportunistic pathogen. *Clin Infect Dis.* 15 Feb 2006, Vol. 42, 4, pp. 451-459.

Skov Sørensen, U B, et al. 2016. Capsular Polysaccharide Expression in Commensal Streptococcus Species: Genetic and Antigenic Similarities to Streptococcus pneumoniae. *mBio.* 15 Nov 2016, Vol. 7, 6, pp. e01844-16.

Smith, M D and Guild, W R. 1979. A plasmid in Streptococcus pneumoniae. *J Bacteriol.* Feb 1979, Vol. 137, 2, pp. 735-739.

Song, J, et al. 2005. SNPsFinder--a web-based application for genome-wide discovery of single nucleotide polymorphisms in microbial genomes. *Bioinformatics.* 1 May 2005, Vol. 21, 9, pp. 2083-2084.

Standish, A J, Stroeher, U H and Paton, J C. 2007. The pneumococcal two-component signal transduction system RR/HK06 regulates CbpA and PspA by two distinct mechanisms. *J Bacteriol.* Aug 2007, Vol. 189, 15, pp. 5591-5600.

—. 2005. The two-component signal transduction system RR06/HK06 regulates expression of cbpA in Streptococcus pneumoniae. *Proc Natl Acad Sci U S A.* 24 May 2005, Vol. 102, 21, pp. 7701-7706.

Starr, C R and Engleberg, N C. 2006. Role of hyaluronidase in subcutaneous spread and growth of group A streptococcus. *Infect Immun.* Jan 2006, Vol. 74, 1, pp. 40-48.

Sternberg, G M. 1885. The pneumonia-coccus of Friedlander (Micrococcus Pasteuri. Sternberg). *Am J Med Sci.* 1885, Vol. 90, pp. 106-123.

Swoboda, J G, et al. 2010. Wall teichoic acid function, biosynthesis, and inhibition. *Chembiochem.* 4 Jan 2010, Vol. 11, 1, pp. 35-45.

Swofford, D. 1999. PAUP*: Phylogenetic analysis using parsimony (* and other methods). 1999.

Tamura, K, et al. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* Aug 2007, Vol. 24, 8, pp. 1596-1599.

Tettelin, H, et al. 2001. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. *Science.* 20 Jul 2001, Vol. 293, 5529, pp. 498-506.

Tettelin, H, et al. 2015. Genomics, genetic variation, and regions of differences. [book auth.] J Brown, S Hammerschmidt and C Orihuela. *Streptococcus pneumoniae: Molecular mechanisms of host-pathogen interactions.* s.l. : Elsevier Science Publishing Co Inc., 2015, 5.

Thompson, J D, Higgins, D G and Gibson, T J. 1994. CLUSTAL W. *Nucleic acids research.* 1994, Vol. 22, 22, pp. 4673–4680.

Todd, E W. 1938. *J. Path. and Bact.* 47, 1938, 423.

Todorova, K. 2010. β-Laktam-Resistenz in Streptococcus spp.: Eine neue Resistenzdeterminante murE. *Dissertation.* 2010.

Todorova, K, et al. 2015. Transfer of penicillin resistance from Streptococcus oralis to Streptococcus pneumoniae identifies murE as resistance determinant. *Mol Microbiol.* Sep 2015, Vol. 97, 5, pp. 866-880.

Tong, H H, et al. 2000. Evaluation of the virulence of a Streptococcus pneumoniae neuraminidase-deficient mutant in nasopharyngeal colonization and development of otitis media in the chinchilla model. *Infect Immun.* Feb 2000, Vol. 68, 2, pp. 921-924.

Treangen, T J and Salzberg, S L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 29 Nov 2011, Vol. 13, 1, pp. 36-46.

Tritt, A, et al. 2012. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One.* 2012, Vol. 7, 9, p. e42304.

Tu, A H, et al. 1999. Pneumococcal surface protein A inhibits complement activation by Streptococcus pneumoniae. *Infect Immun.* Sep 1999, Vol. 67, 9, pp. 4720-4724.

van Tonder, A J, et al. 2014. Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Comput Biol.* 21 Aug 2014, Vol. 10, 8, p. e1003788.

Watson, D A, et al. 1993. A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America.* 1993, Vol. 17, 5, pp. 913–924.

Weber, B, et al. 2000. The fib locus in Streptococcus pneumoniae is required for peptidoglycan crosslinking and PBP-mediated beta-lactam resistance. *FEMS Microbiol Lett.* 1 Jul 2000, Vol. 188, 1, pp. 81-85.

Weiser, J N, et al. 1994. Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonization. *Infect Immun.* Jun 1994, Vol. 62, 6, pp. 2582-2589.

Weiser, J N, Ferreira, D M and Paton, J C. 2018. Streptococcus pneumoniae: transmission, colonization and invasion. *Nat Rev Microbiol.* Jun 2018, Vol. 16, 6, pp. 355-367.

Westerik, N, et al. 2016. Novel Production Protocol for Small-scale Manufacture of Probiotic Fermented Foods. *J Vis Exp.* 10 Sep 2016, 115.

Whatmore, A M, et al. 2000. Genetic relationships between clinical isolates of Streptococcus pneumoniae, Streptococcus oralis, and Streptococcus mitis: characterization of "Atypical" pneumococci and organisms allied to S. mitis harboring S. pneumoniae virulence factor-encoding genes. *Infect Immun.* Mar 2000, Vol. 68, 3, pp. 1374-1382.

Winslow, C E A, et al. 1920. The families and genera of the bacteria: final report of the committee of the Society of American Bacteriologists on characterization and classification of bacterial types. *J Bacteriol.* 1920, 5, pp. 191-229.

Woese, C R. 2000. Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A.* 18 Jul 2000, Vol. 97, 15, pp. 8392-8396.

Wyres, K L, et al. 2012. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. *Genome Biol.* 16 Nov 2012, Vol. 13, 11, p. R103.

Zähner, D, et al. 2011. Mitis Group Streptococci Express Variable Pilus Islet 2 Pili. *PLoS ONE.* 2011, Vol. 6, 9.

Zhang, J R, et al. 2000. The polymeric immunoglobulin receptor translocates pneumococci across human nasopharyngeal epithelial cells. *Cell.* 15 Sep 2000, Vol. 102, 6, pp. 827-837.

Zhang, Q, et al. 2015. Streptococcus pneumoniae genome-wide identification and characterization of BOX element-binding domains. *Mol Inform.* Nov 2015, Vol. 34, 11-12, pp. 742-752.

Acknowledgements

All work included in this thesis was done in the department of Microbiology of the University of Kaiserslautern under the guidance of Prof. Dr. Regine Hakenbeck.

I would like to thank my colleagues of the Department of Microbiology, Dalia Denapaite, Reinhold Brückner, Patrick Maurer, Irma Ochigava, Katya Todorova, Tina Becker and Yvonne Schähle for their work on the presented publications and assistance during my time at the department. Naturally, I am also grateful to my colleagues Katharina Peters, Inga Schweizer, Marina Meiers, Julia Sauerbier.

Moreover, I would like to thank Prof. Dr. Reinhold Brückner of the microbiology department of the University of Kaiserslautern for data and words of advice regarding small RNAs as well as Dr. Dalia Denapaite for her advice and support.

Further thanks to my new colleague Jochen Scammell, his father-in-law Philip Scammell and their family for help in English language.

Most thanks to my wife Ekaterine Toklikishvili, who supported me during this thesis and, together with our children Uwe Tamazi and Toma Peter, endured my frequent absence during this time. You gave me the strength to complete this work.

Finally, I would like to thank Prof. Dr. Regine Hakenbeck of the microbiology department of the University of Kaiserslautern for her guidance during my work, support and words of advice as well as for her unswerving patience.

Appendices

The appendices listed below are provided on the CD ROM.

1. The entire PhD thesis in PDF format.

2. The supplementary (Fig. S1) material of the paper: “Rieger, M, Mauch, H and Hakenbeck, R. 2017. Long persistence of a *Streptococcus pneumoniae* 23F clone in a cystic fibrosis patient. mSphere. 7 Jun 2017, Vol. 2, 3, pp. e00201-17”

3. The supplementary (Fig. S1-3; Table S1-4) material of the paper: “Denapaite D, Rieger M, Köndgen S, Brückner R, Ochigava I, Kappeler P, Mätz-Rensing K, Leendertz F, Hakenbeck R. Highly variable *Streptococcus oralis* strains are common among viridans Streptococci isolated from primates. *mSphere.* 9 Mar 2016, Vol. 1, 2, pp. e00041-15”

4. The supplementary material of the unpublished work:

    Table S1: Read statistics of sequenced bacterial strains.

    Table S2: Comparison of D122 and D141 genomes – genes containing SNVs.

    Table S3: SNVs in all three ST10523 genomes.

    Table S4: Comparison of Hu15 and Hu17 genomes – Regions of divergent sequence.

    Table S5: Proteins of Hu15 and Hu17 not present in Hu19A.

    Table S6: Comparison of Hu15 and Hu17 genomes – genes containing SNVs.

    Table S7: Assembly statistics of streptococci obtained from human and primate.

    Table S8: Presence of *S. oralis* Uo5 proteins in *S. oralis* strains obtained from primates and humans.

    Table S9: Complete *S. pneumoniae* genomes used for determination of *S. pneumoniae* R6-specific proteins.

    Table S10: *S. pneumoniae* R6-specific proteins at intraspecies comparison.

    Table S11: Presence of *S. pneumoniae* R6 proteins in other pneumococci and representatives of other species.

    Table S12: Proteins of *S. pneumoniae* R6 not found in compared genomes.

5. The sequence comparison and analysis tool with simple manual.

Curriculum Vitae

Personal information

Name	Martin Rieger
Nationality	German

Education

1992 – 1994	Goethe-Gymnasium, Bensheim/Bergstraße
1994 – 2001	Tilemann-Schule, Limburg
2002 – 2008	University of applied sciences Gießen-Friedberg, Gießen Diploma thesis: “MARSlab-Software-Suite; Software-Paket für die Submittierung, Archivierung und Analysevorverarbeitung auf Basis der MARS-Datenbank zum Zwecke der Transkriptomanalyse”
2012 –	PhD studies at the department of microbiology, University of Kaiserslautern

Experience

09/2001 – 09/2002	Military service, Hessenkaserne/Stadtallendorf
06/2009 – 07/2012	Technical assistant at the department of microbiology, University of Kaiserslautern
08/2012 – 12/2014	PhD student at the department of microbiology, University of Kaiserslautern
02/2016 –	IT consultant and software developer at the Neox AG für Informationstechnologie, Pirmasens

Hiermit versichere ich, die vorliegende Dissertation in der Abteilung Mikrobiologie der Universität Kaiserslautern selbstständig durchgeführt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben.

Kaiserslautern, Juni 2020 Martin Rieger